Министерство образования Республики Беларусь

Учреждение образования «Белорусский государственный университет информатики и радиоэлектроники»

Национальная академия наук Беларуси

Объединенный институт проблем информатики

DHTechnologies & Data Nubes (Austin, USA)

Middlesex University (London, UK)

ООО «АктивХост РУ» (Москва, Россия)

ООО «Бел Хуавэй Тэхнолоджис»

ИООО «ЭПАМ СИСТЕМЗ»

BEZNext (Chicago, USA)

Invisi BV (Netherlands)

IBM (NY, USA)

Third International Conference and Expo BIG DATA and ADVANCED ANALYTICS

May 3-4, 2017

Minsk, Belarus

СБОРНИК МАТЕРИАЛОВ
III МЕЖДУНАРОДНОЙ НАУЧНО-ПРАКТИЧЕСКОЙ КОНФЕРЕНЦИИ

Минск БГУИР, 2017

УДК 004.6(082)
ББК 32.973.26-018.2я43
Б59

BIG DATA and Advanced Analytics: collection of materials of the third international scientific and practical conference. (Minsk, Belarus, May 3 – 4, 2017) / editorial board: M. Batura [etc.]. – Minsk, BSUIR, 2017. – 350 с.

Б 59

ISBN 978-985-534-323-2

В сборнике опубликованы результаты научных исследований и разработок в области BIG DATA and Advanced Analytics для оптимизации IT-решений, бизнес-решений, а также представлены технологии и инструментарий для аналитики в области медицины, образования и других тематических исследований.

УДК 004.6(082)
ББК 32.973.26-018.2я43

ISBN 978-985-534-323-2

ОРГАНИЗАЦИОННЫЙ КОМИТЕТ КОНФЕРЕНЦИИ

Председатель, Батура М.П.

Ректор Белорусского государственного университета информатики и радиоэлектроники, доктор технических наук, профессор, академик «Международной академии наук высшей школы», заслуженный работник образования Республики Беларусь, Республика Беларусь

Сопредседатель, Boris Zibitsker

PhD, President and CEO BEZNext, Chicago, USA
Emeritus professor of BSUIR

Сопредседатель, Dominique A. Heger

PhD, President and CEO DHTechnologies & Data Nubes, Austin, USA

Заместитель председателя, Дик С.К.

Первый проректор БГУИР, кандидат физико-математических наук, доцент, Республика Беларусь

Члены организационного комитета

James Ogunleye, PhD, Professor at Middlesex University and the Editor of the International Journal of Developments in Big Data and Analytics, UK

Alain Biem, PhD, Opera Systems, formerly at IBM Watson Research – Big Data, NY, USA

Dirk Marc Guy Stroo, PhD, Owner of Invisi, Netherlands, Owner of Act On Insight, Belarus, Information Innovation Leader, Business Intelligence Consultant: Royal Agio Cigars, City of Rotterdam, Nuon

Никульшин Б.В., проректор по учебной работе Белорусского государственного университета информатики и радиоэлектроники, кандидат технических наук, доцент, Республика Беларусь

Лихачевский Д.В., декан факультета компьютерного проектирования Белорусского государственного университета информатики и радиоэлектроники, кандидат технических наук, Республика Беларусь

Nedim Karagenç, MD PhD, Professor at Pamukkale University, medical Faculty, department of medical biology, Turkey

M. Bülent Ozdemir, PhD, Professor at Pamukkale University, medical Faculty, department of anatomi, Turkey

Hakan Akça PhD, Professor. Pamukkale University, Medical Faculty, Department of Medical Biology, Turkey

Цырельчук И.Н., декан факультета непрерывного и дистанционного обучения, заведующий кафедрой проектирования информационно-компьютерных систем БГУИР, кандидат технических наук, доцент, Республика Беларусь

Яшин К.Д., заведующий кафедрой инженерной психологии и эргономики Белорусского государственного университета информатики и радиоэлектроники, кандидат технических наук, доцент, Республика Беларусь

ОРГАНИЗАТОРЫ КОНФЕРЕНЦИИ

Министерство образования Республики Беларусь

Учреждение образования
«Белорусский государственный университет информатики и радиоэлектроники»

Национальная академия наук Беларуси
Объединенный институт проблем информатики

BEZNext (Chicago, USA)

DHTechnologies & Data Nubes (Austin, USA)

ИООО «ЭПАМ СИСТЕМЗ»

ООО «Бел Хуавэй Тэхнолоджис»

IBM (NY, USA)

Middlesex University (London, UK)

ООО «АктивХост РУ» (Москва, Россия)

Invisi BV (Netherlands)

СПОНСОРЫ КОНФЕРЕНЦИИ

GODEL TECHNOLOGIES

sam solutions

itransition SOFTWARE DEVELOPMENT COMPANY

ARTEZIO

ALTOROS™

dev by

Paralect

МОЛОДЁЖНОЕ ТЕЛЕВИДЕНИЕ БГУИР

ИП Тимохов

JET BI

KIAT™

DIGITAL SYSTEMS LABORATORY

Active TECH

ScienceSoft Professional Software Development

СПОНСОРЫ КОНФЕРЕНЦИИ

ООО «Нэкстсофт»

ООО "СЛС ИТ"

brimit

SoverSys
Ваша безопасность и комфорт

ПРИВЕТСТВИЕ УЧАСТНИКАМ КОНФЕРЕНЦИИ

БАТУРА Михаил Павлович
Ректор Белорусского государственного университета информатики и радиоэлектроники, доктор технических наук, профессор

ДИК Сергей Константинович
Первый проректор БГУИР, кандидат физико-математических наук, доцент

Boris ZIBITSKER,
PhD, President and CEO BEZNext, Chicago, USA,
Emeritus professor of BSUIR

Цуй ЦИМИН
Чрезвычайный и Полномочный Посол Китайской Народной Республики в Республике Беларусь

Роберт РАЙЛИ
Временный поверенный в делах США в Беларуси

Али Осам Абед Али
Временный поверенный в делах Республики Ирак в Республике Беларусь

КАРПЕНКО Игорь Васильевич
Министр образования Республики Беларусь

ШЕДКО Дмитрий Геннадьевич
Первый заместитель Министра связи и информатизации Республики Беларусь

ЯНЧЕВСКИЙ Всеволод Вячеславович
Директор Государственного учреждения «Администрация Парка высоких технологий»

МАРТИНКЕВИЧ Александр Михайлович,
Заместитель директора по маркетингу и развитию Государственного учреждения «Администрация Парка высоких технологий»

БОГУШ Вадим Анатольевич
Первый заместитель Министра образования Республики Беларусь, доктор технических наук, профессор

ТУЗИКОВ Александр Васильевич
Генеральный директор Объединенного института проблем информатики Национальной академии наук Беларуси, доктор физико-математических наук, профессор, член-корреспондент, Национальной академии наук Беларуси

ДЮБКОВ Владимир Константинович
Ректор института ИТ и бизнес-администрирования IBA, кандидат технических наук, доцент

ДЕМИЧЕВ Алексей Евгеньевич
Генеральный директор ЗАО «Итранзишэн»

О Г Л А В Л Е Н И Е

ПЛЕНАРНЫЕ ДОКЛАДЫ

INCORPORATION OF PRESCRIPTIVE ANALYTICS FOR PERFORMANCE ENGINEERING AND DYNAMIC PERFORMANCE MANAGEMENT OF BIG DATA APPLICATIONS

B. ZIBITSKER, PhD
President and CEO
BEZNext, Emeritus professor of BSUIR

CEO BEZNext, Chicago, USA
E-mail: bzibitsker@beznext.com

Abstract: In a complex Big Data environment applications compete for resources and affect each other performance. Selection of Machine Learning Algorithms and Machine Learning Libraries and Big Data YARN's Scheduler, Queues and Containers rules can significantly affect accuracy, performance and scalability of Big Data applications.

We will review how predictive and prescriptive analytics can be used to develop recommendations optimizing selection of the algorithms, priorities, concurrency and resource allocation to effectively satisfy Service Level Goals for each of the workloads.

During this presentation we will review how prescriptive analytics can be used for evaluation of the options and justification of the changes necessary to proactively and continuously meeting Service Level Goals (SLGs) for each workload. We will review how comparison of the actual results with expected are used to organize a feedback control.

BIG DATA ANALYTICS - MISSING OR MESSY DATA, WHAT NOW?

D. A. HEGER, PhD
CEO/Owner DHTechnologies
& Data Nubes

DHTechnologies, Austin, TX

Introduction. Missing data scenarios are common Big Data problems in domains such as biology, finance, medicine, life-science, research, or climatic science (to name a few). They can arise from different sources such as mishandling of samples, low signal-to-noise ratios, measurement errors, hardware failures, transmission errors, no-response, or deleted aberrant values. Rubin introduced the notion of the distribution of missingness as a way to classify the conditions under which missing data should be treated. Little and Rubin distinguish among data missing completely at random (MCAR), data missing at random (MAR), and data missing not at random (MNAR):

1. Data missing completely at random (MCAR) describes scenarios where the probability of an instance (case) having a missing value for a variable does not depend on either the known values or the missing data, respectively. An example would be a sensor outage that results into missing some measurements.

2. Data missing at random (MAR) describes scenarios where the probability of an instance having a missing value for a variable may depend on the known values but not on the value of the missing data itself. For example, suppose men are less likely (compared to women) to respond to an income question in a survey, but the likelihood of responding is independent of their actual income. In this case, unbiased gender-specific income estimates can be made if one has data on the gender variable (as an example by replacing the missing income value with the gender-specific median income).

3. Data missing not at random (MNAR) describes scenarios where the probability of an instance having a missing value for a variable may depend on the value of that variable. For example, this may occur if either low or high income subjects (or both) are less likely to answer the income question in a survey. This is the most complex type of missing data and in many cases, there is no good value to substitute for the missing one. But, by just dropping these subjects, the results will be biased and hence such an approach is normally not suggested.

It has to be pointed out that missing data introduces an element of ambiguity into the data analysis cycle. Missing data can affect properties of statistical estimators such as the mean, variance, or percentage, resulting in a loss of information power and misleading conclusions. A variety of techniques have been proposed for substituting missing values with statistical predictions, a process that is generally referred to as missing data imputation. To reiterate, it is very often the case that the weakest link in a Big Data analytics study is the quality of the available data sets. It is a fact that the more one can find out about why the data sets are as they are, the more one can develop a case on the pattern of the missing data, as well as on a rationale on why the pattern may or may not matter. It has to be pointed out though that in many studies, just eliminating missing cases prior to the analysis is not viewed as a legitimate solution to the problem.

Some of the rather standard techniques to address missing data (such as list-wise deletion, single

mean imputation, or single regression imputation) may lead to biased estimates of the model parameters. To illustrate, a simple mean substitution method leaves the mean unchanged but decreases the variance! Some of the more appropriate techniques (in many cases) for dealing with missing data are the multiple imputation (MI) and the (full information) maximum likelihood (ML) estimation that incorporate the missing data points in the analysis (see Enders and Enders & Peugh). MI involves imputing a range of random plausible values for missing data, a process that results in several complete data sets that can then be analyzed. One of the advantages of this approach is that the technique introduces variability into the distribution of cases with missing data (simulating the messy world we are living in), a process that is more likely to represent the population than imputing a single value for each missing case. Partial data actually contributes to the estimation of the model's parameters by implying probable values for missing scores via the correlations among the variables. Expectation maximization (EM), a common method for obtaining ML estimates with incomplete data sets, treats the model parameters (rather than the data points themselves) as missing values to be estimated and borrows information from the existing data at successive iterations until differences between successive iterations are minor. Missing data can pose a number of additional problems in multilevel data structures, depending on the sampling design underlying the data set, the extent to which the data are missing at each level, and whether or not the data can be assumed to be missing at random. In some modeling situations, there may be a considerable amount of missing data. Compared with single-level analyses, the difficulties presented by multilevel analyses scenarios concerns the likelihood that the missing data at one level (Level 2) may be linked to the missing data at Level 1.

In any Big Data analytics project, it is paramount to distinguish between data preparation and data analysis. While preparing the data for analysis, it is imperative to determine the amount of missing data, as well as the missing data pattern(s). It is essential to note that the reality is that there is a very small chance to get the missing (real) data back in the first place. Hence, a data scientist always has to deal with the problem of missing information. The quality of the analysis though depends on assumptions one makes on the pattern of the missing data and what is reasonable to conclude about those patterns. Now, what one can do about the missing data becomes the pressing concern!

Definition - Missing Completely at Random (MCAR). Suppose the probability of an observation being missing does not depend on observed or unobserved measurements. In mathematical terms, one can stipulate:

$$\Pr(\mathbf{r} \mid \mathbf{y_o}, \mathbf{y_m}) = \Pr(\mathbf{r}) \tag{1}$$

Then one can state that the observation is missing completely at random (MCAR). Note that in a sample survey, MCAR may be labeled as uniform non-response. If data sets are MCAR, then consistent results with missing data can be obtained by performing the analyses as if there were no missing data. But, doing so will result in some loss of information. So that implies that with MCAR data sets, the analysis only provides valid inferences with complete data sets! So does an analysis based on moment based estimators (for example generalized estimating equations) and other estimators derived from consistent estimating equations. The term consistent estimating equations refers to functions of the data and unknown parameters whose expectation (taken over the complete data at the population parameter values) is 0. Under MCAR, they still have expectation zero and so still lead to valid inferences.

Stating the same mathematically, an estimating equation can be written as $U\ (y, \theta)$ and at the estimate $\hat{\theta}, U\ (y, \hat{\theta}) = 0$. The estimating equation is consistent because $EU\ (Y, \theta) = 0$ (where θ reflects the population parameter value). It remains consistent if the data reflects missing completely at random (MCAR) because $EU\ (Y_o, \theta) = 0$. A simple example of a consistent estimating equation is the sample mean $U\ (y, \theta) = \bar{y} - \theta$. With MCAR a single imputation or a multiple imputation (MI) method can be considered. It has to be pointed out that with a single imputation method, it may be

difficult to generate valid variance estimates though! The author of this report always suggest to at least consider an MI approach with MCAR.

Definition - Missing At Random (MAR). After considering MCAR, a second scenario may come up. That is, what are the most general conditions under which a valid analysis can be done using only the observed data and no information about the missing value mechanism, Pr(r | yo, ym)? The answer here is when, given the observed data, the missingness mechanism does not depend on the unobserved data. Mathematically stated:

$$\Pr(r \mid y_o, y_m) = \Pr(r \mid y_o). \text{ (MAR)} \quad (2)$$

This is equivalent to stating that the behavior of 2 runs who share observed values have the same statistical behavior on the other observations, whether observed or not. To illustrate:

Table 1. Some Measurements

Measurements	*Features*					
Collection Run	*1*	*2*	*3*	*4*	*5*	*6*
1	**1**	**3**	**4.3**	**3.5**	**1**	**4.6**
2	**1**	**3**	**NA**	**3.5**	**NA**	**NA**

As runs 1 and 2 in Table 1 have the same values (where values for both runs are available), given these observed values (under MAR), features 3, 5 and 6 from run 2 have the same distribution (not the same value!) as features 3, 5 and 6 from run 1. It has to be pointed out that under MAR, the probability of a value being missing will generally depend on observed values, so it does not correspond to the intuitive notion of random. The important idea is that the missing value mechanism can be expressed solely in terms of observations that are observed. Unfortunately, this scenario can hardly ever be definitively determined from the data at hand! An example of a MAR mechanism may be the scenario where 2 measurements of the same variable are made concurrently. If they differ by more than a given amount, a 3d measurement is taken. This 3d measurement is missing for those scenarios that do not differ by the given amount specified for the 1st 2 measurements.

A special case of MAR is known as uniform non-response within classes. To illustrate, one seeks to collect data on income and property tax bands. Typically, those with higher incomes may be less willing to reveal the actual numbers. So a simple average of incomes from people who actually responded will be downwardly biased. However, assuming one has everyone's property tax band and given that the property tax band non-response to the income question is random, then the income data is missing at random. The reason, or mechanism, for it being missing depends on the property band. Given the property band, missingness does not depend on income itself. Therefore, to get an unbiased estimate of income, one has to first average the observed income within each property band. As data is missing at random given a property band, these estimates will be valid. To get an estimate of the overall income, one has to combine these estimates, weighted by the proportion in each property band.

It further has to be pointed out that likelihood methods (such as EM) are valid for MAR as well. However, in general, non-likelihood methods (based on simple completers, moments, estimating equations) are not valid for MAR per se. So ordinary means and other simple summary statistics that are based on the observed data will be biased. Finally, with likelihood, the term ignorable is often used to refer to a MAR mechanism. But it is the mechanism (the model for $\Pr(R \mid y_o)$) that is ignorable, not the missing data! To summarize, MAR scenarios can be analyzed via multiple imputation (MI) methods or likelihood-based methods (such as EM).

Definition - Missing Not At Random (MNAR). When neither MCAR nor MAR holds, one states that the data sets are *Missing Not At Random* (MNAR). Now, in a likelihood setting, the *missingness*

mechanism is labeled *non-ignorable*. All this basically implies that while even accounting for all the available observed information, the reason for observations being missing *still depends on the unseen observations themselves*.

In other words, the probability of a data value being missing is related to the unobserved values. To illustrate, a study may focus on analyzing tumors. So that study requires measurements of the actual tumor sizes. The data may show that smaller tumors are less like to have sizes recorded (maybe due to any detection delays). So it is harder to actually measure the size of smaller tumors. Hence, while there may be good data available for larger tumor sizes, the sizing data for smaller tumors may be missing. So with MNAR, the important question to be answered is how is the data related to the unobserved value?

Similar to MAR, MNAR scenarios can be analyzed via Multiple Imputation (MI) methods or likelihood-based methods (such as EM). It has to be pointed out that MNAR is much more complex to deal with and basically requires modeling the process yielding the missing values.

Definition - Multiple Imputation/ Multiple imputation refers to a statistical technique for analyzing incomplete data sets (data sets for which some entries are missing). The actual application of the technique requires 3 steps, imputation, analysis, and pooling.

1. Imputation (impute = fill in). Impute the missing data m times to produce m complete data sets. Regression models, Bayesian ideas (with MCMC), or Fully Conditional Specifications (that works well for categorical data) can be used here.

2. Analysis. Analyze each data set by using a standard statistical procedure. This step results in m analyses.

3. Pooling. Integrate the m analysis results into a final result See Rubin or Schafer.

Note: Rubin has shown that if the method to create imputations is proper, then the resulting inferences will be statistically valid. The real challenge here is in the imputation phase (aka the construction of the m completed data sets). This step accounts for the process that created the missing data. A typical problem here could be that the missing data is actually related to its value (aka wealthy people tend to skip income questions in surveys). It further has to be reemphasized that missing entries

can appear anywhere in the data and that the method used in the imputation step must foresee the intended complete-data analyses. But, the repeated analysis step on the imputed data is actually somewhat simpler than the same analysis without imputation, as there is no need to bother with the missing data per se. The pooling step consists of computing the mean over the (m) repeated analysis, its variance, and its confidence interval or p value. In general, these computation are reasonably simple. Some common data imputation techniques are (to just name a few):

- MI Mean
- K-nearest neighbors (KNN)
- fuzzy K-means (FKM)
- singular value decomposition (SVD)
- Bayesian principal component analysis (BPCA)
- Bayesian ideas (regression, MCMC)
- Multiple imputations by chained equations (MICE)
- Fully conditional specifications (FCS)

To further discuss MI, The MCMC and the FCS methods are elaborated on in more detail here. Both methods are very popular iterative methods for performing multiple imputations for missing data patterns. The Markov Chain Monte Carlo (MCMC) method is widely used for Bayesian inference (Schafer) and is considered as one of the most popular iterative algorithm for multiple imputation scenarios. The basic flow is that one commences with some reasonable starting values for the mean, variance, and the covariance among a given set of variables. Next, one divides the sample into subsamples, each having the same missing data pattern (the same set of variables present and missing). For each missing data pattern, one uses the starting values to construct linear regressions for imputing the missing data, using all the observed variables in that pattern as predictors. One then imputes the

missing values, making random draws from the simulated error distribution, which results into a single completed data set.

Using this data set with missing data imputed, one recalculates the mean, variance and covariance and then makes a random draw from the posterior distribution of these parameters. Finally, one uses these drawn parameter values to update the linear regression equations needed for imputation. This process is typically repeated many times. For example, one may run 200 iterations of the algorithm before selecting the first completed data set, and then allow for another 100 iterations between each successive data set. So producing 5 data sets (as an example) requires 600 iterations (each of which generates a data set). Why so many iterations? The first 200 (burn-in) iterations are designed to ensure that the algorithm has converged to the correct posterior distribution. Then, allowing 100 iterations between successive data sets gives the confidence that the imputed values in the different data sets are statistically independent. If all assumptions are satisfied, the MCMC method produces parameter estimates that are consistent, asymptotically normal, and almost fully efficient. Full efficiency would require an infinite number of data sets, but a relatively small number normally gets one very close.

An alternative algorithm is known as the fully conditional specification (FCS) or multiple imputation by chained equations (MICE) (Brand, Van Buuren, Oudshoorn). This method is attractive because of its ability to impute both quantitative and categorical variables appropriately. It allows one to specify a regression equation for imputing each variable with missing data (usually linear regression for quantitative variables and logistic regression (binary, ordinal, or unordered multinomial) for categorical variables. Under logistic imputation, imputed values for categorical variables will also be categorical. Imputation proceeds sequentially, usually starting from the variable with the least missing data and progressing to the variable with the most missing data. At each step, random draws are made from both the posterior distribution of the parameters and the posterior distribution of the missing values. Imputed values at one step are used as predictors in the imputation equations at subsequent steps (something that never happens in MCMC algorithms). Once all missing values have been imputed, several iterations of the process are repeated before selecting a completed data set. Although attractive, FCS has 2 major disadvantages (compared to the linear MCMC method). First, it is much slower, computationally. Second, FCS itself has no theoretical justification. By contrast, if all assumptions are met, MCMC is guaranteed to converge to the correct posterior distribution of the missing values. FCS carries no such guarantee, although simulation results by Van Buuren are very encouraging.

Summary - Multiple Imputation (MI). Just as the single imputation methods, multiple imputation fills in estimates for the missing data. But to capture the uncertainty in those estimates, MI estimates the values multiple (m) times. As MI utilizes an imputation method that has an error term built in, the multiple estimates should be similar, but not identical. The result basically is multiple data sets (m) with identical values for all of the non-missing values and slightly different values for the imputed values in each data set. The statistical analysis of interest (such as logistic regression) is performed separately on each data set and the results are then consolidated. Because of the variation in the imputed values, there should also be variation in the parameter estimates, leading to appropriate estimates of standard errors and p-values, respectively.

Definition - Maximum Likelihood (ML). Depending on the pattern and the amount of missing data, a potentially legit approach may be to analyze the full, incomplete data set via a maximum likelihood estimation. This method does not impute any data, but rather uses each cases available data to compute maximum likelihood estimates. The maximum likelihood estimate of a parameter equals to the value of the parameter that is most likely to have resulted in the observed data.

When data points are missing, one can factor the likelihood function. The likelihood is computed separately for those cases with complete data on some variables and those with complete data on all variables. These 2 likelihoods are then maximized together to find the estimates. Like multiple imputation, this method provides unbiased parameter estimates and standard errors. One advantage

of ML is that it does not require the careful selection of variables used to impute values on (necessary with MI).

Some Actual Guidelines

In the next few paragraphs, some basic guidelines are provided. It has to be pointed out though that if more than 5% to 10% of the data is missing, a scenario like that is considered a potential major source of a serious bias that has to be addressed accordingly.

– Assuming that the proportion of missing data is ≤ 0.05. If less than 5% of the data is missing, studies have shown that it matters little what imputation method is chosen or whether one adjusts the variance of the regression coefficient estimates for having imputed data in this case. So for continuous variables, imputation via the median value should be adequate. For categorical predictors, the most frequent category can be used. A complete case analysis may be an option here as well.

– Assuming that the proportion of missing data is between **0.05 and 0.15**. In this scenario, if a predictor is unrelated to all of the other predictors, imputations can be done the same way as above (impute a reasonable constant value). If the predictor is correlated with other predictors, develop a customized model (such as via a transcan function - a nonlinear additive transformation and imputation function) to foresee the predictor from all of the other predictors. Then impute missing data with predicted values. For categorical variables, classification trees are good methods for developing customized imputation models. For continuous variables, ordinary regression can be used if the variable in question does not require a non-monotonic transformation to be predicted from the other variables. For either the related or unrelated predictor case, variances may need to be adjusted for imputation. The author of this report suggests multiple imputation or maximum likelihood methods here while single imputation methods may be considered.

– Assuming that the proportion of missing data is **> 0.15**. Such a scenario requires the same considerations as in the previous case and adjusting variances for imputation is even more important. To estimate the strength of the effect of a predictor that is frequently missing, it may be necessary to refit the model on the subject of observations for which that predictor is not missing. In this case either multiple imputation or maximum likelihood methods are preferred for most models.

Summary. Analyzing and cleansing (missing) data is paramount in order to achieve actual, accurate conclusions. To illustrate, while using the same ANN to conduct a regression study, using 2 differently cleansed data-sets as the input, there is a high probability that the 2 different input sets will result into 2 different answers/conclusions. Further, in any data analysis project, greater than 5% to 10% of missing data points is considered a potential source of a very serious bias condition. It is always important to consider and scrutinize the model/environment that produces the missing data.

Many studies have shown that in most scenarios, either a multiple imputation or an ML method does provide excellent results (even for MCAR). Further, if the missing data reflects MNAR, there is a need to consider and scrutinize the model that actually gives rise to the missing data. Plus, if the missingness is strongly related to the value of the variable, the problem becomes rather complex (a fact that cannot just be ignored).

References

[1]. Allison, P. (2002). Missing data. Thousand Oaks, CA: Sage.

[2]. Bodner, T. E. (2008).What improves with missing data imputations? Structural Equation Modeling, 15, 651-675.

[3]. Enders, C. K. (2011). Missing not at random models for latent growth curve analyses. Psychological Methods, 16(1), 1-16.

[4]. Enders, C.K., & Peugh, J. L. (2004). Using an EM covariance matrix to estimate structural equation models with missing data: Choosing an adjusted sample size to improve the accuracy of inferences. Structural Equation Modeling, 11, 1–19.

[5]. Hedeker, D., & Gibbons, R. D. (1997). Application of random-effects pattern-mixture models for missing data in longitudinal studies. Psychological Methods, 2, 64–78.

[6]. Hox, J. (2010). Multilevel applications: Techniques and applications (2nd Edition). New York:

Routledge.

[7]. Kish, L. (1989). Statistical design for research. New York: Wiley.

[8]. Larsen, R. (2011). Missing data imputation versus full information maximum likelihood with second-level dependencies. Structural Equation Modeling, 18(4), 649-662.

[9]. Little, R.J.A & Rubin, D. B. (2002). Statistical analysis with missing data. New York, NY: Wiley.

[10]. Rubin, D.B. (1976). Inference and missing data. Biometrika, 63, 581–592.

[11]. Peugh, J.L. & Enders, C.K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. Review of Educational Research, 74, 525-556.

[12]. Schafer, J. (2005, November). Missing data in longitudinal studies. A review. Paper presented at the Annual Meeting of the American Association of Pharmaceutical Scientists, Nashville, TN.

[13]. Chandola T, Brunner E, Marmot M. (2006). Chronic stress at work and the metabolic syndrome: prospective study. BMJ 332:521-5. PMID:16428252

[14]. Greenland S, Finkle WD. (1995). A critical look at methods for handling missing covariates in epidemiologic regression analysis. Am J Epidemiol 142(12):1255-64.

[15]. Fleiss JL, Levin B, Paik MC. (2003). Statistical Methods for Rates and Proportions, 3rd ed. Hoboken NJ, John Wiley & Sons.

[16]. Harrell Jr FE. (2001). Regression Modeling Strategies With Applications to Linear Models, Logistic Regression, and Survival Analysis. New York, Springer-Verlag.

[17]. Huberman M, Langholz B. (1999). Application of the missing-indicator method in matched case-control studies with incomplete data. Am J Epidemiol 150(12):1340-5.

[18]. Li X, Song X, Gray RH. (2004). Comparison of the missing-indicator method and conditional logistic regression in 1:m matched case-control studies with missing exposure values. Am J Epidemiol 159(6):603-610.

[19]. Moons KG, Grobbee DE. (2002). Diagnostic studies as multivariable, prediction research. J Epidemiol Community Health 56(5):337-8.

[20]. Roth, P. (1994). Missing data: A conceptual review for applied psychologists. Personnel Psychology 47:537-560.

[21]. Royston P. (2004). Multiple imputation of missing values. The Stata Journal 4(3):227-241.

[22]. Royston P. (2005a). Multiple imputation of missing values: update. The Stata Journal 5(2):188-201.

[23]. Royston P. (2005b). Multiple imputation of missing values: update of ice. The Stata Journal 5(4):527-536.

[24]. Royston P. (2007). Multiple imputation of missing vlaues: further update of ice, with an emphasis on interval censoring. The Stata Journal 7(4):445-464.

[25]. Schnonlau M. (2006). Stata software package, hotdeckvar.pkg, for hotdeck imputation. http://www.schonlau.net/stata/.

[26]. Steyerberg EW. (2009). Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. New York, Springer.

[27]. Twisk JWR. (2003). Applied Longitudinal Data Analysis for Epidemiology: A Practical Guide. Cambridge, Cambridge University Press.

[28]. van Buuren S, Boshuizen HC, Knook DL. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. Statistics in Medicine 18: 681–694.

[29]. Vandenbrouchke JP, von Elm E, Altman DG, et al. (2007). Strengthening and reporting of observational studies in epidemiology (STROBE): explanation and elaboration. Ann Intern Med 147(8):W-163 to W-194.

[30]. Brand, J.P.L. (1999) Development, Implementation and Evaluation of Multiple Imputation Strategies for the Statistical Analysis of Incomplete Data Sets. Dissertation, Erasmus University,Rotterdam.

[31]. Cornwell, C. and Rupert, P., "Efficient Estimation with Panel Data: An Empirical Comparison of Instrumental Variable Estimators," Journal of Applied Econometrics, 3, 1988, pp. 149-155.

[32]. Dempster, A. P., Nan M. Laird and Donald B. Rubin. 1977. "Maximum Likelihood Estimation from Incomplete Data Via the EM Algorithm." Journal of the Royal Statistical Society, Series B 39: 1-38.

[33]. Heckman, James J. 1976. "The Common Structure of Statistical Models of Truncated, Sample Selection and Limited Dependent variables, and a Simple Estimator of Such Models." Annals of Economic and Social Measurement 5:475-492.

[34]. Little, Roderick J. A. and Donald B. Rubin. 1987. Statistical Analysis with Missing Data. New York: Wiley. SAS Global Forum 2012 Statistics and Data Analysis

[35]. Molenberghs, G. and Kenward, M.G. (2007) Missing Data in Clinical Studies. Chichester, UK: John

Wiley and Sons Ltd.

[36]. Mroz, T. A. (1987) "The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions." Econometrica 55, 765–799.

[37]. Raghunathan, Trivellore E., James M. Lepkowski, John Van Hoewyk and Peter Solenberger (2001) "A multivariate technique for multiply imputing missing valuesusing a sequence of regression models." Survey Methodology, 27:85-95.

[38]. Rubin, Donald B. 1976. "Inference and Missing Data." Biometrika 63: 581-592.

[39]. Rubin, Donald B. 1987. Multiple Imputation for Nonresponse in Surveys. New York: Wiley.

[40]. Schafer, Joseph L. 1997. Analysis of Incomplete Multivariate Data. London: Chapman & Hall.

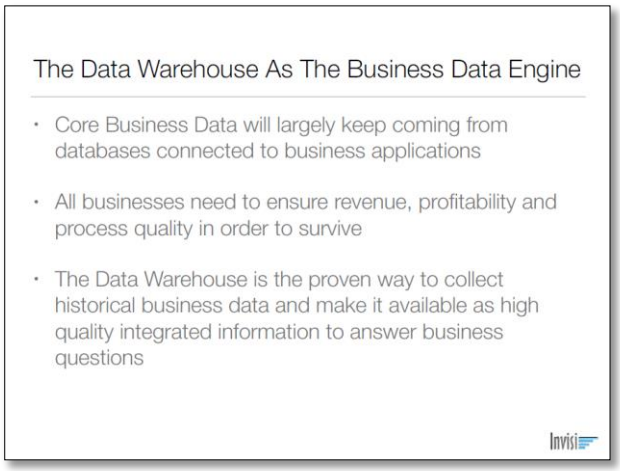
[41]. Van Buuren, S., J.P.L. Brand, C.G.M. Groothuis-Oudshoorn and D.B. Rubin (2006) "Fully conditional specification in multivariate imputation." Journal of Statistical Computation and Simulation 76: 1046-1064.

[42]. Van Buuren, S., and C.G.M. Oudshoorn (2000). Multivariate Imputation by Chained Equations: MICE V1.0 User's Manual. Report PG/VGZ/00.038. Leiden: TNO Preventie en Gezondheid.

BIG BELARUS DATA. THE YEAR OF TRUTH FOR BELARUS, BIG DATA AND ANALYTICS

M. G. STROO, PhD
Owner of Invisi, Netherlands, Owner of Act On Insight, Belarus, Information Innovation Leader, Business Intelligence Consultant: Royal Agio Cigars, City of Rotterdam, Nuon

Invisi BV, Netherlands



Analytics Maturity - Crawl Before You Run

- People first learn to crawl, then to walk, then to run
- In analytics maturity, first learn to see what happened, then why did it happen, then what may happen:
  - Reporting and scorecarding
  - Dashboarding and OLAP
  - Forecasting and Predictive Analytics
- First look inwards to your structured process data, then deal with larger amounts of more varied customer and market data

Data Science As Business Innovation Driver

- Companies also possess more and more customer data
- This customer data can be analysed for patterns and trends
- Once Data Mining, now Data Science, is how data is analysed for valuable insights
- Open source tools and software languages like R and Python bring Data Science within reach of many
- This may lead to product and process innovation even for smaller companies

Innovative Business Models to Turn Data Into Profit

- Data generated by your business processes can be used as new products or services
- This data goes from a byproduct of the process to the centre of a new product
- The new products can open up new markets and new clients
- These products typically follow one of a couple of basic patterns

Data Driven Product Patterns

- Sell data
- Innovate products through data
- Swap commodity offerings into value-added services
- Create interaction in the value chain
- Create a network of value based on data exchange

Sell Data Example

- Bank analyses account transactions with specific retailers
- Bank anonymises this data
- Bank benchmarks this data with selected competitors
- Bank sells the result to these specific retailers

Product Innovation Example

- Bank offers checking accounts to consumers
- Bank offers household expenses dashboard to their consumers based on their account transactions
- Bank offers this dashboard also to non-customers who can upload data from their own bank
- Bank offers recommendations based on comparisons, such as energy spending compared to similar households

Data Driven Issues To Consider

- These issues may have limited impact on the original process, but big impact on your data driven product:
  - Data quality
  - Data continuity
- Interference with other processes

Invisi

The State of Big Data and Analytics in Belarus

- Large software services companies in Belarus with data related professionals largely ignore Belarus market
- Middle and large private Belarus companies are seen as having low budgets and limited interest for analytics
- State controlled Belarus companies are seen as not interested in analytics, business performance management and not thinking value driven

Invisi

Overcoming Belarus BI Challenges

- Increase numbers and level of data related and analytics professionals in Belarus
- Reduce knowledge drain of high potentials to other countries
- Use Agile project methods to rapidly implement BI in Belarus state and private companies
- Use innovative methods and tools to decrease development time of data warehouses
- Implement standard and conformed logical data models for core business processes for fast insight in areas like finance, sales and production

Invisi

Overcoming Belarus BI Challenges

- Use open source and low cost tooling as backbone of an analytics architecture
  - Database
    - PostgreSQL, MariaDB, MongoDB
  - ETL and data warehouse automation
    - Pentaho, Talend, Quipu
  - BI server, visualisation and OLAP
    - Pentaho, Jaspersoft, Palo
- Improve on existing open source tooling and create new ones from joint Belarus university and enterprise research

Invisi

Belarus Data Innovation Opportunities

- Reserve budget for data related research and innovation
- Review current internal business and customer data
- Start data warehouse projects in Belarus companies
- Start exploring innovation based on existing company and customer data
- Start Smart City initiatives in major Belarus cities

Invisi

Belarus Data Research Opportunities

- Effectiveness of various visualisation options
- Data virtualisation
- Data Warehouse automation
- Data Lake and Data Warehouse architectures
- Hadoop based Data Warehousing

Invisi


Belarus Data Challenge
Bring Business Intelligence and Data Warehousing to one hundred medium sized Belarus companies per year
Bring Data Science initiatives to ten larger Belarus companies per year
Create hundreds of new Data related Belarus jobs per year
Transform three Belarus companies to data centred companies per year

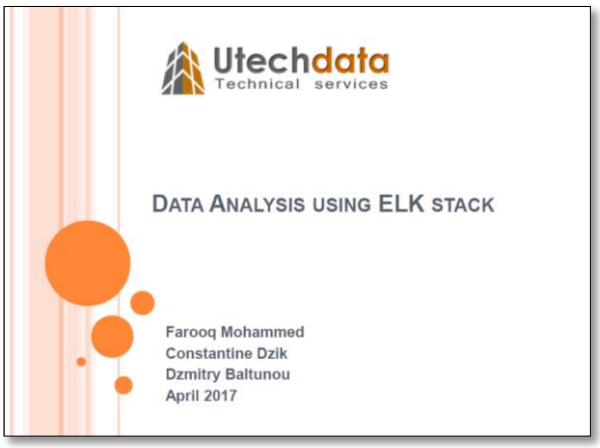
DATA ANALYSIS USING ELK STACK

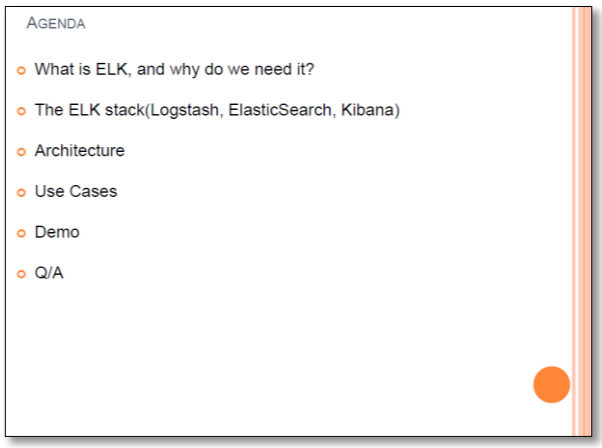
F. MOHAMMED
Utech LLC location – Madison

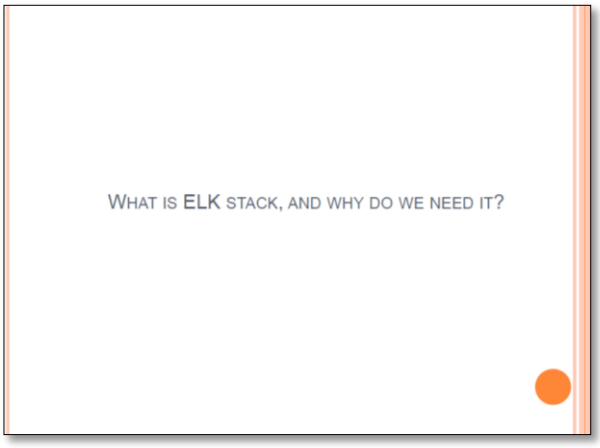
D. BALTUNOU
Utech LLC location – Madison

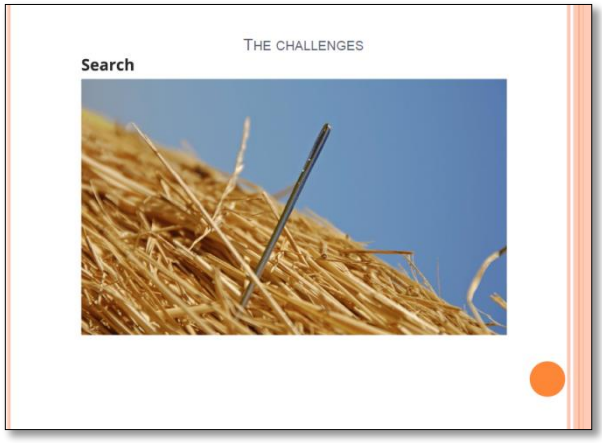
C.S. DZIK
Utech LLC location – Madison

Utech Solution Inc, Madison, USA
E-mail: constantine.dzik@utechdata.com









ONE SOLUTION: ELK

What is the ELK stack

- Elasticsearch
  - Search server
  - Based on Apache Lucene
- Logstash
  - Data pipeline
  - Processes logs and other data
  - Plugins
- Kibana
  - Web frontend for Elasticsearch

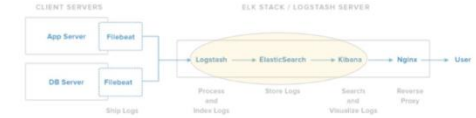
ELK Overview

Elasticsearch: NoSQL DB Storage

Logstash: Data Collection & Digestion

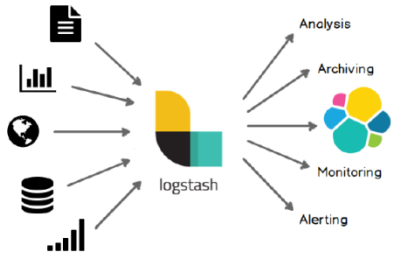
Kibana: Visualization standard.

Stack Components by Type
- Shippers
- Brokers
- Storage / Processing
- Visualization

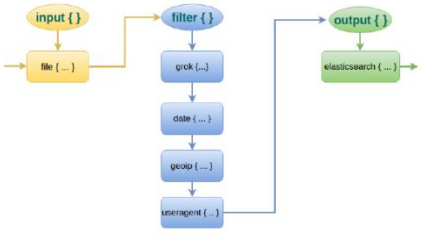


THE ELK STACK(LOGSTASH, ELASTICSEARCH, KIBANA)

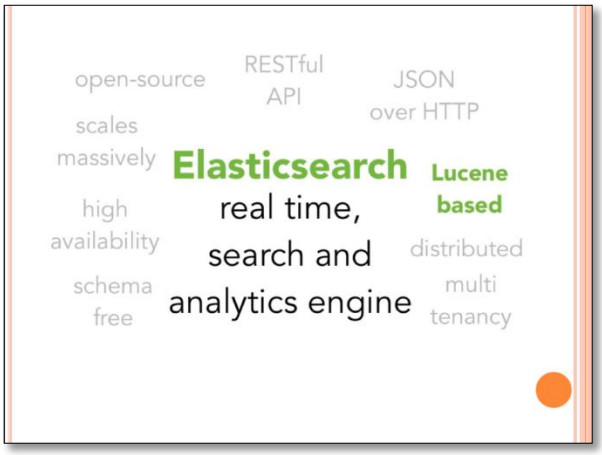
LOGSTASH



LOGSTASH CONFIG STRUCTURE



LOGSTASH CONFIG FILE EXAMPLE

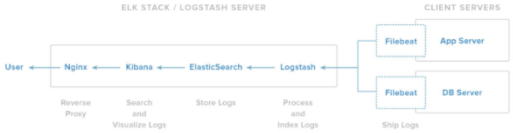


KIBANA

- Kibana is an open source analytics and visualization platform designed to work with Elasticsearch
- Use Kibana to search, view, and interact with data stored in Elasticsearch indices.
- Easily perform advanced data analysis and visualize your data in a variety of charts, tables, and maps.
- Easy to understand large volumes of data. Its simple, browser-based interface enables you to quickly create and share dynamic dashboards that display changes to Elasticsearch queries in real time.

ARCHITECTURE

Our ELK stack setup has four main components:
- **Logstash**: The server component of Logstash that processes incoming logs
- **Elasticsearch**: Stores all of the logs
- **Kibana**: Web interface for searching and visualizing logs, which will be proxied through Nginx
- **Filebeat**: Installed on client servers that will send their logs to Logstash, Filebeat serves as a log shipping agent that utilizes the *lumberjack* networking protocol to communicate with Logstash



USE CASES

USAA

USAA reduces security incidents by searching 3-4 billion security events a day, running Python scripts, building custom applications to mine the data, and utilizing Watcher, the Elasticsearch alerting and notification extension.

FICO

FICO is leveraging Elasticsearch, unstructured and semistructured data to significantly improve the performance of FICO's analytics models. FICO's Analytic Modeler for Text products makes these advanced analytics and visualizations available to end users in an intuitive and interactive way. FICO has integrated advanced descriptive, diagnostic, and predictive analytics with Elasticsearch, and extended Kibana to provide advanced visualizations against same.

NETFLIX

Netflix

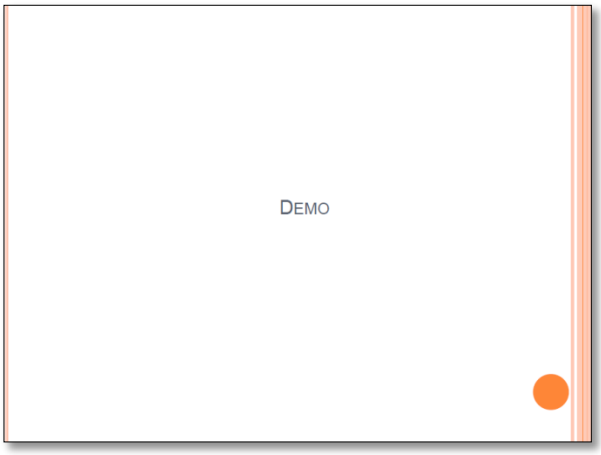
Netflix messages millions of customers a day across many channels – email, push notifications, text, voice calls, etc – via its messaging platform: a distributed system made up of a series of separate applications. They use Elasticsearch for higher message deliverability and operational excellence.

GitHub

GitHub uses Elasticsearch to continually index the data from an ever-growing store of over 8 million code repositories, comprising over 2 billion documents. Using Elasticsearch, GitHub was able to let users easily search this data. One goal of GitHub's Elasticsearch implementation is to index everything that is publicly available on GitHub.com and make it easy to find. Full-text searching is fully supported, but searching based on a wide variety of criteria is also possible and dead simple.
On GitHub you can search for a project that uses Clojure as the primary language, and has had activity over the past month, and all this functionality is powered by Elasticsearch.

Facebook

Facebook has been using Elasticsearch for 3 plus years, having gone from a simple enterprise search to over 40 tools across multiple clusters with 60+ million queries a day and growing.


DEMO


Q&A


OUR CONTACTS
Utech LLC
102 Mason Cir
Madison, MS
39102, USA
Tel. + 1-847-868-2292
www.utechdata.com
emails:
mohammed.farooq@utechdata.com
dzmitry.baltunou@utechdata.com
constantine.dzik@utechdata.com

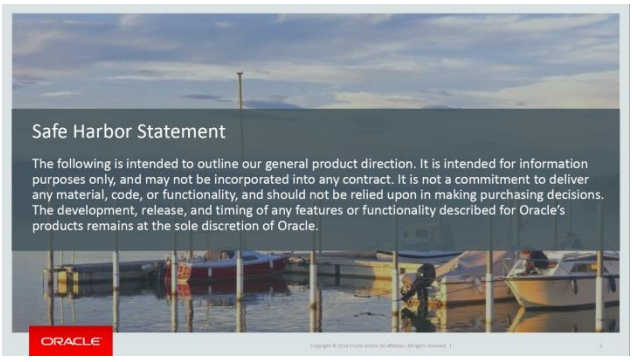

REFERENCES
1 https://www.elastic.co
2 https://www.digitalocean.com/
3 https://www.elastic.co/use-cases
4 https://www.oreilly.com/learning/a-guide-to-elasticsearch-5-and-the-elkelastic-stack

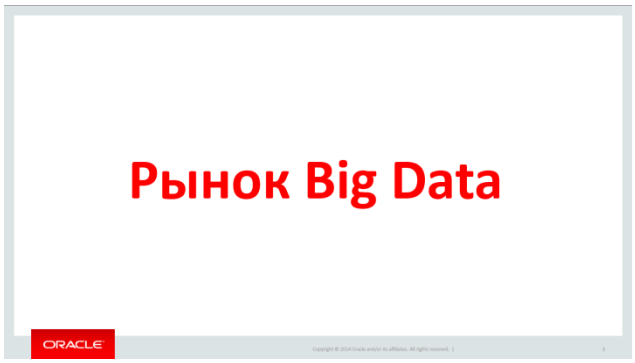
ORACLE BIG DATA: A MUST-HAVE TECHNOLOGY IN 2017

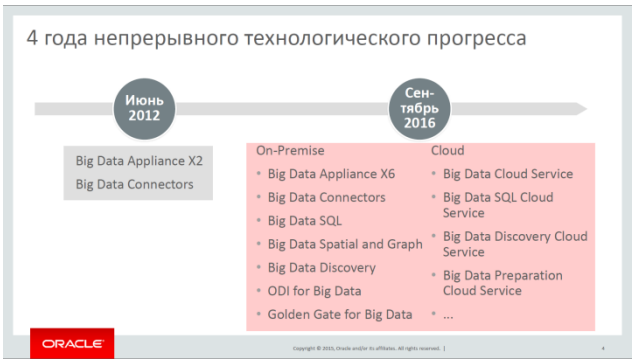
N. USPENSKIY
BigData Specialist, ORACLE Representative in Russia, Kazakhstan, Mongolia, Russia

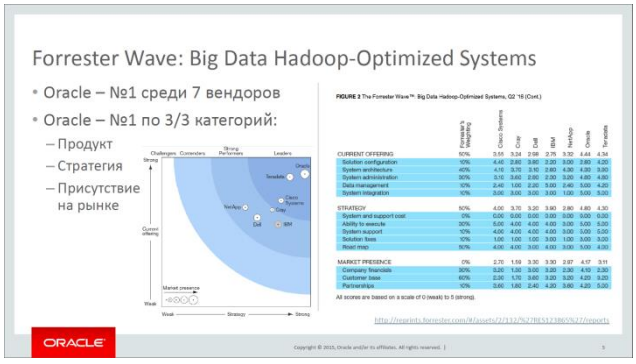
ORACLE, Russia
E-mail: nikita.uspenskiy@oracle.com

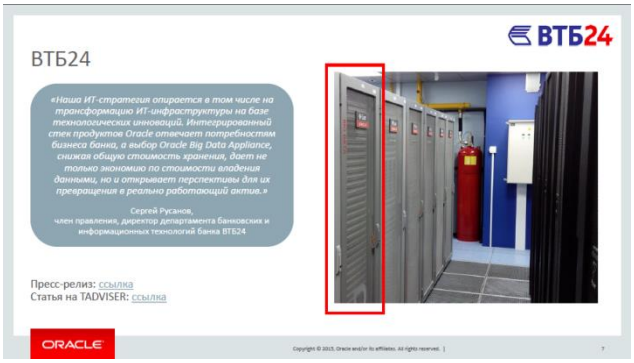
Abstract. The speaker will share information major Big Data market trends and most popular use cases in CIS region. The presentation will also cover Oracle Big Data portfolio overview and recommendations for Big Data projects implementation.



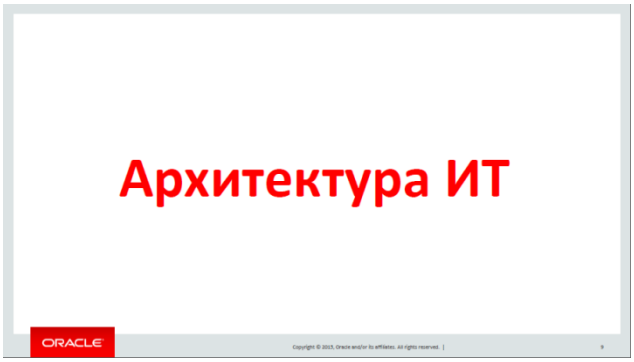




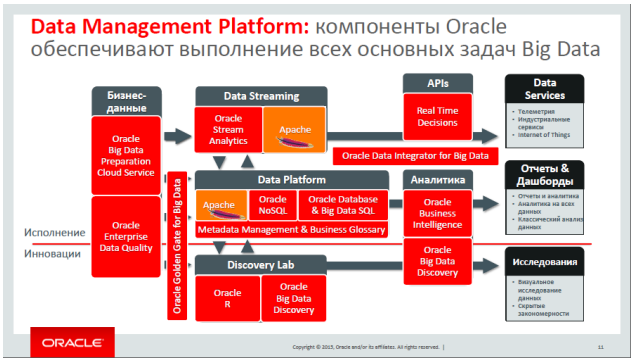


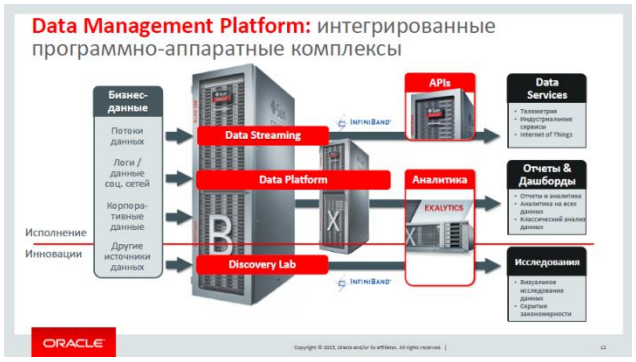

ВТБ24
ВТБ24
«Наша ИТ-стратегия опирается в том числе на трансформацию ИТ-инфраструктуры на базе технологических инноваций. Интегрированный стек продуктов Oracle отвечает потребностям бизнеса банка, а выбор Oracle Big Data Appliance, снижая общую стоимость хранения, дает не только экономию по стоимости владения данными, но и открывает перспективы для их превращения в реально работающий актив.»
Сергей Русанов,
член правления, директор департамента банковских и информационных технологий банка ВТБ24
Пресс-релиз: ссылка
Статья на TADVISER: ссылка
ORACLE

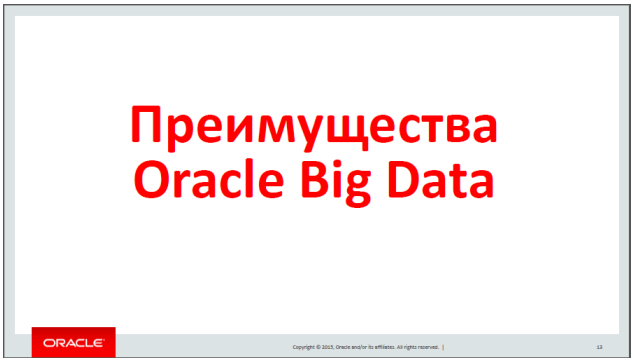

Альфа Банк
Альфа-Банк
«В начале мы хотели увидеть, действительно ли анализ больших данных приносит те плоды, о которых красиво рассказывают менеджеры по продажам. Мир становится другим, объем данных растет. Анализ информации становится источником дополнительного дохода для бизнеса. Мы в начале пути. У нас есть большие планы по использованию внешних и внутренних данных. Важны не просто технологии, а важен комплекс – соединение технологии, людей, которые это делают и конкретных целей бизнеса»
Ирина Елистратова, главный директор BI
Описание проекта на сайте Cnews: ссылка
Статья на сайте Банковского обозрения: ссылка
ORACLE

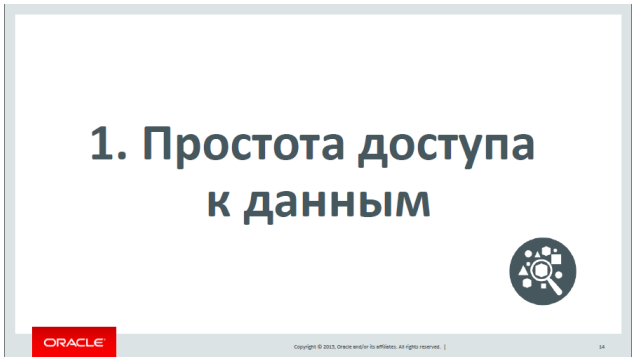

Архитектура ИТ
ORACLE

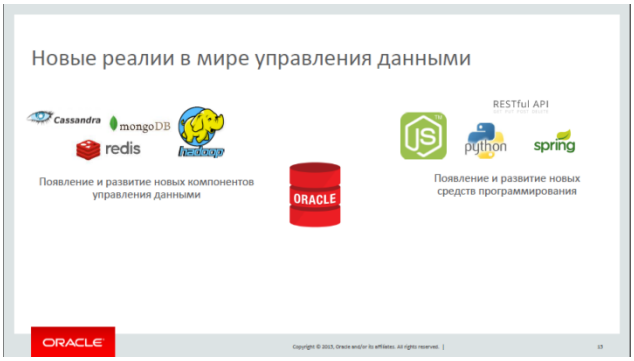

Data Management Platform: общая архитектура
Бизнес-данные
Потоки данных
Логи / данные соц. сетей
Корпора-тивные данные
Другие источники данных
Исполнение
Инновации
Data Streaming
Data Platform
Reservoir
Data Factory
Warehouse
Discovery Lab
APIs
Аналитика
Data Services
Отчеты & Дашборды
Исследования
ORACLE


Data Management Platform: компоненты Oracle обеспечивают выполнение всех основных задач Big Data
Бизнес-данные
Oracle Big Data Preparation Cloud Service
Oracle Enterprise Data Quality
Oracle Golden Gate for Big Data
Исполнение
Инновации
Data Streaming
Oracle Stream Analytics
Apache
APIs
Real Time Decisions
Oracle Data Integrator for Big Data
Data Platform
Apache
Oracle NoSQL
Oracle Database & Big Data SQL
Metadata Management & Business Glossary
Аналитика
Oracle Business Intelligence
Discovery Lab
Oracle R
Oracle Big Data Discovery
Data Services
Отчеты & Дашборды
Исследования
ORACLE


Data Management Platform: интегрированные программно-аппаратные комплексы
Бизнес-данные
Потоки данных
Логи / данные соц. сетей
Корпора-тивные данные
Другие источники данных
Исполнение
Инновации
Data Streaming
Data Platform
Discovery Lab
INFINIBAND
APIs
Аналитика
EXALYTICS
Data Services
Отчеты & Дашборды
Исследования
ORACLE

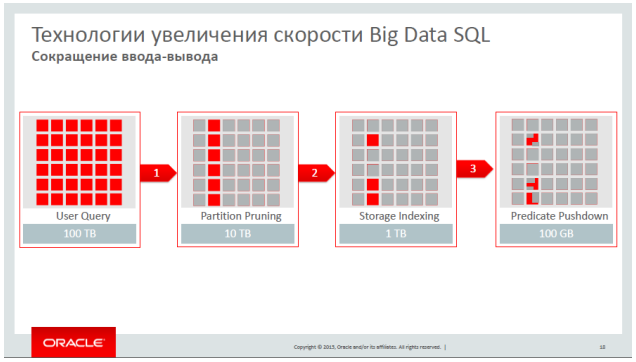

Преимущества Oracle Big Data
ORACLE

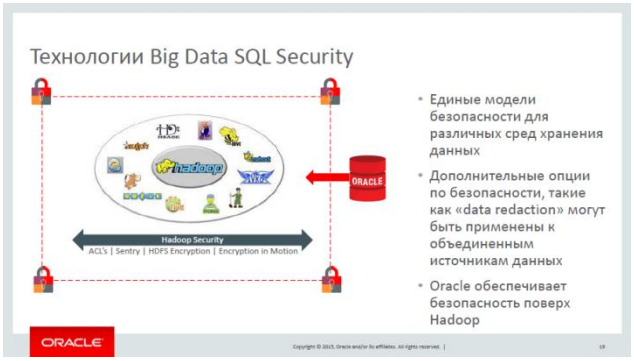

1. Простота доступа к данным
ORACLE

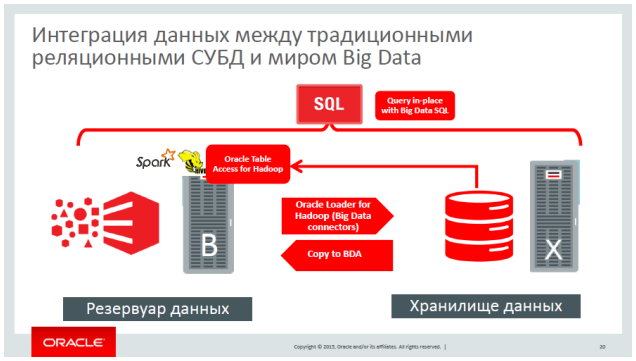

Новые реалии в мире управления данными
Cassandra
mongoDB
redis
hadoop
Появление и развитие новых компонентов управления данными
ORACLE
RESTful API
node
python
spring
Появление и развитие новых средств программирования
ORACLE

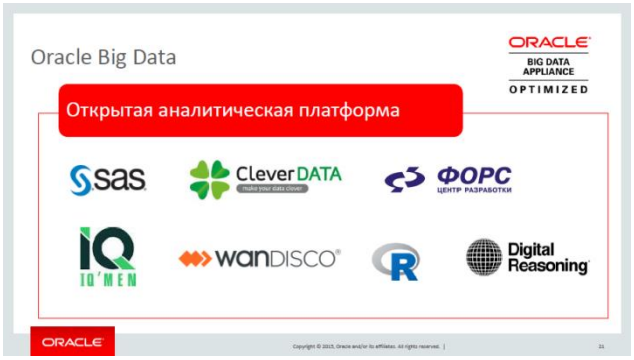

Big Data SQL
SQL
Hadoop
Логи
NoSQL
Профили клиентов
Relational
Транзакции
Информационная безопасность:
ВСЕ данные в рамках единых общих политик обеспечения безопасности
ORACLE

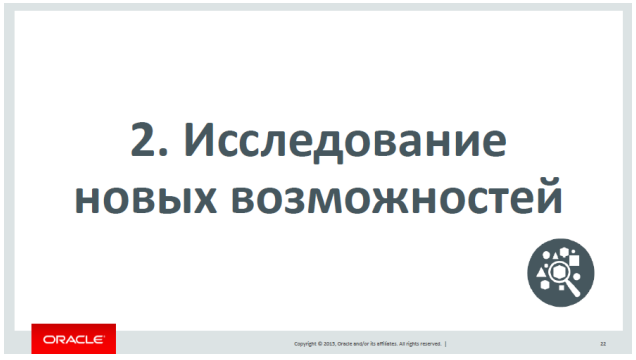

Единый интерфейс обращения к данным
Oracle Big Data SQL - нет необходимости переписывать существующие приложения
SQL
SQL
Быстрый возврат выборочных данных
Предобработка запроса на узлах кластера
Часть данных
Предобработка запроса на серверах хранения Exadata
B
X
Hadoop и NoSQL
Oracle Database 12c
ORACLE


Технологии увеличения скорости Big Data SQL
Сокращение ввода-вывода
User Query
100 TB
1
Partition Pruning
10 TB
2
Storage Indexing
1 TB
3
Predicate Pushdown
100 GB
ORACLE


Технологии Big Data SQL Security
hadoop
ORACLE
Hadoop Security
ACL's | Sentry | HDFS Encryption | Encryption in Motion
Единые модели безопасности для различных сред хранения данных
Дополнительные опции по безопасности, такие как «data redaction» могут быть применены к объединенным источникам данных
Oracle обеспечивает безопасность поверх Hadoop
ORACLE


Интеграция данных между традиционными реляционными СУБД и миром Big Data
SQL
Query in-place with Big Data SQL
Spark
Oracle Table Access for Hadoop
Oracle Loader for Hadoop (Big Data connectors)
Copy to BDA
B
X
Резервуар данных
Хранилище данных
ORACLE


Oracle Big Data
ORACLE
BIG DATA APPLIANCE
OPTIMIZED
Открытая аналитическая платформа
sas
CleverDATA
ФОРС
ЦЕНТР РАЗРАБОТКИ
IQ MEN
wanDISCO
R
Digital Reasoning
ORACLE


2. Исследование новых возможностей
ORACLE

Из данных в Hadoop не так легко извлечь **пользу**

- Существующие аналитические средства не подходят
  - Требуется много усилий для подготовки данных
  - Ручная обработка наборов данных
  - Очистка данных и их подготовка сильно зависят от ETL инструментов
  - Аналитические запросы должны быть определены заранее
- Нет единого инструмента для анализа
  - Несогласованность данных и визуализации
  - Нет единого «источника правды»
- Нужен полный функционал
- Сложные компоненты Hadoop
  - Pig, Oozie, Sqoop, Hive, Spark
- Сложно найти специалистов
  - Языки программирования (например, Map Reduce, Python, Scala)
  - Мат. статистика и машинное обучение
  - Интерфейсы с командной строкой

Аналитика больших данных требует новых подходов

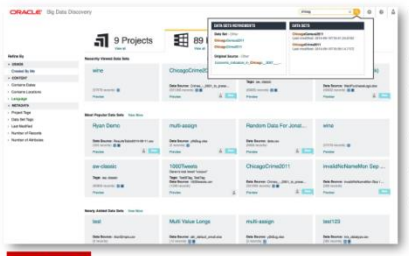
Интуитивный, интерактивный и **наглядный** пользовательский интерфейс

понятный **всем**, для быстрого **поиска, изучения, трансформации** и **анализа** данных в Hadoop,

а также для **общего пользования** результатами

Big Data Discovery. **Визуальный интерфейс к Hadoop**

Легко **находите** нужные наборы данных

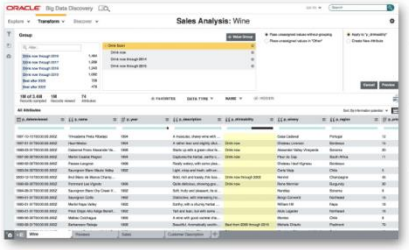


- Доступ к интерактивному **каталогу** данных, расположенных в Hadoop
- Удобный **поиск** и **навигация** в наборах данных
- Аннотации, краткие описания и **рекомендации** к наборам данных
- **Загрузка** собственных данных
- Автоматическое **обогащение** данных
- Навигация по собственным и коллективным **проектам**

Изучайте данные и раскрывайте их потенциал

- Понимание **структуры** данных. Визуализация атрибутов по типам
- **Сортировка атрибутов по степени их важности** на основе алгоритмов машинного обучения
- Распределение данных, **качество данных** и выбросы
- Понимание **корелляции** между атрибутами
- **Оценка целесообразности** дальнейшей работы с набором данных

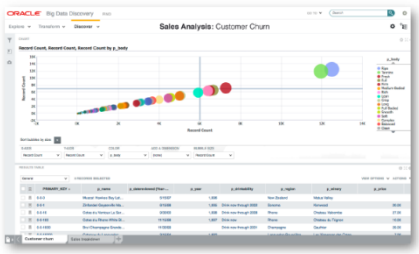
Трансформируйте и обогащайте данные



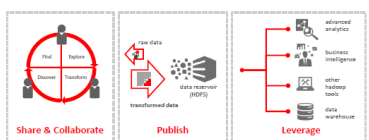
- Интуитивный интерфейс для **трансформации** данных
- Расширяемая **библиотека функций** (замена значений, конвертация типов, группировки, слияния и т.д.)
- **Предварительный просмотр** результатов, отмена, подтверждение и повторное выполнение трансформаций
- Тестирование на небольшом наборе данных в оперативной памяти, применение на полном наборе данных в **Hadoop**

Исследуйте данные для получения новых знаний

- **Объединение** различных данных для более глубокого анализа
- Фильтрация данных с помощью мощного **поисковика** и интуитивной **навигации с подсказками**
- Интерфейс **«Drag & drop»** для создания интерактивных визуальных дэшбордов
- **Публикация** в Hadoop отчетов на основе смешанных наборов данных



Делитесь результатами с Вашими коллегами



Oracle Big Data Discovery отлично работает со всех **экосистемой Big Data**

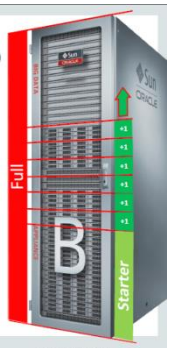
- **Делитесь и взаимодействуйте** с Вашей командой
  - Делитесь проектами, закладками, снапшотами и т.д.
- **Публикуйте** данные в Hadoop
  - Трансформации и обогащения могут быть применены к исходным наборам данных
  - Публикация подготовленных наборов данных в HDFS
- **Используйте** результаты для работы с другими инструментами
  - Публикуйте данные в Hadoop в форматах, оптимизированных для аналитических инструментов (например, ORAAH)
  - Используйте стандартные компоненты Haddop (например, Pig, Hive, Impala, Python, т.д.)

3. Развертывание, сопровождение и TCO

ORACLE

BIG DATA APPLIANCE – аппаратная платформа

- **МОЩНЫЙ**
  - До **792 ядер** с **4.6TB RAM** (расширяется до 13.8TB)
  - До **1728 TB** дискового пространства
  - 40 Gb/s InfiniBand
- **МОДУЛЬНЫЙ**
  - Минимум 6 узлов и расширение по **96TB**
  - до 20+ систем в кластере
- **ОПТИМИЗИРОВАННЫЙ**
  - Оптимизации для **Linux**
  - Сконфигурированная **Java**
  - Сетевые оптимизации
  - Оптимизации для **Hadoop**
  - Единая точка обновления и патчирования
- **БЕЗОПАСНЫЙ**
  - Kerberos аутентификация
  - Авторизация Apache Sentry
  - Аудит Oracle Audit Vault
  - Шифрование данных
  - Шифрование трафика
- **УПРАВЛЯЕМЫЙ**
  - Централизованное управление через OEM 12c



ORACLE

BIG DATA APPLIANCE – программное обеспечение

ORACLE

- Oracle Linux 5 или Oracle Linux 6
- Oracle Java – JDK 8
- Oracle R Distribution
- Oracle Big Data Appliance Enterprise Manager Plug-In
- **Big Data SQL (опционально, лицензируется отдельно)**
- **Big Data Discovery (опционально, лицензируется отдельно)**
- **Oracle NoSQL Database Enterprise Edition (опционально, лицензируется отдельно)**

cloudera

- Cloudera Enterprise 5 – Data Hub Edition
- Cloudera's Distribution including Apache Hadoop (CDH) with support for YARN and MR2
- Cloudera Impala
- HBase (as well as support for Accumulo)
- Cloudera Search
- Apache Spark
- Cloudera Manager including:
- Cloudera Back-up and Disaster Recovery (BDR)
- Cloudera Navigator

ORACLE

Эластичное масштабирование от Starter Rack to Multi-Rack

- **Начальная конфигурация включает 6 BDA узлов и все свитчи**
  - **Можно добавлять по одному BDA узлу**
- **Системы предыдущих поколений можно расширять новыми узлами**

ORACLE

Что вам НЕ придется делать с Big Data Appliance (1/2)

- Подбирать и оптимизировать компоненты – серверы, диски, количество дисков, процессоры, сети, память и т.п.
- Заключать отдельный договор о поддержке с Cloudera
- Собирать кластер
- Настраивать сетевые коммутаторы
- Инсталлировать операционную систему на каждом узле и
- Отслеживать и устанавливать оптимальные версии драйверов и прошивок для каждого компонента
- Настраивать операционную систему для оптимальной производительности (у нас же очень много данных!)
- Настраивать Java

ORACLE

Что вам НЕ придется делать с Big Data Appliance (2/2)

- Инсталлировать дополнительное ПО от Cloudera
- Тестировать работоспособность и производительность каждого узла кластера
- Заниматься самостоятельно трудоемкой процедурой многуровнего апгрейда и патчирования BIOS, OS, Java, Hadoop и т.п.
- И просто следить за тем, что нужно что-то проапгрейдить
- Изучать как это все сделать без остановки и прерывания работы пользователей
- Заниматься дизайном перебалансировки кластера при его расширении
- И т.д. и т.п.

ORACLE

TCO и развитие системы

- Экономия на лицензировании:
  - По кол-ву дисков (например, Oracle Big Data SQL):
    - Больше размер дисков → меньше лицензий для нужного объема данных
- Экономия на тех. поддержке ПО:
  - 12% от стоимости Oracle Big Data Appliance, тех. поддержка Cloudera учтена
  - Напрямую покупать дороже (цена на узел $3000-$7000, узлы могут быть «слабее»)
- Западные заказчики переходят на Enterprise-решения
- Рынок развивается в сторону конвергентных систем, а не подхода «сделай сам»

ORACLE

Техническая поддержка

- Premier Support: 24x7, русский язык в рабочие часы с 9:00 до 18:00
- Покрывает весь «стек» программно-аппаратного комплекса:
  - Аппаратные компоненты
  - Системное ПО (ОС, Java и т.д.)
  - Платформенное ПО Cloudera
- Единый интерфейс обращения – через Oracle
- 12% от стоимости решения в год

ORACLE

4. Производительность

1DMP (CleverDATA)

Тестирование партнерского решения в РФ

- Результаты тестирования:
  - Обработка более 101,5 тысяч (в 1,45 раз больше изначально ожидаемого числа) запросов в секунду всего на одном узле Oracle Big Data Appliance. Время отклика на запрос в 99% случаев не превысило 1,17 миллисекунды (в 1,7 раза лучше ожидаемого).
  - Время классификации Интернет-страниц и построение пользовательских профилей на 6 узлах Oracle Big Data Appliance составило 11 минут 17 секунд и оказалось в 5,43 раза лучше ожидаемых результатов.

http://cleverdata.ru/1dmp-na-oracle-big-data-appliance/

«Число клиентов Биржи данных растет, а это значит, что актуальными становятся вопросы увеличения производительности и масштабируемости решения. Мы решили эти задачи за счет портирования платформы на Oracle Big Data Appliance, что позволит клиентам 1DMP строить на облачной площадке собственные решения по управлению большими данными, накапливать и за короткое время классифицировать большие объемы информации, формировать профили клиентов для целей таргетированного маркетинга. Комплекс Oracle Big Data Appliance позволил нам обеспечивать короткое минимальное время отклика, что является критичным на данном рынке работы с данными»

Денис Афанасьев, генеральный директор CleverDATA

IQPLATFORM® (Айкумен ИБС)

Тестирование партнерского решения в РФ

- Результаты тестирования:
  - Скорость поиска в архиве выросла в 9 раз – с 40 000 до 366 600 документов в секунду

http://www.iqmen.ru/iqmen-oracle_exastack

«Ускорение обработки данных с применением программно-аппаратных комплексов Oracle Exadata, Oracle Exalogic и Oracle Big Data Appliance позволяет нашим клиентам наиболее эффективно обрабатывать большие массивы накопленной информации, используя быстрый онлайн-доступ к необходимым данным для их комплексного анализа прямо на входном потоке. Таким образом, открываются возможности решения нового класса исследовательских задач, требующих обработки данных в режиме реального времени»

Дмитрий Часовской,
руководитель проектного отдела «Айкумен ИБС»

ForSMedia (ФОРС – ЦР)

Тестирование партнерского решения в РФ

- Результаты тестирования:
  - Скорость поиска в архиве выросла в 9 раз – с 40 000 до 366 600 документов в секунду

http://www.fors.ru/pressroom/news/1892/

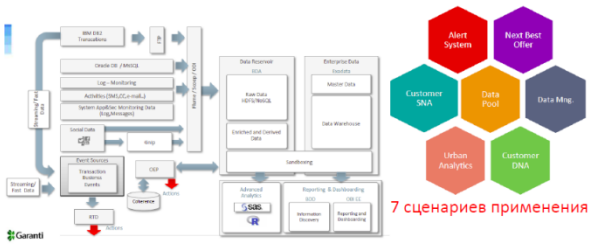
«Столь впечатляющие результаты тестирования ForSMedia на Oracle Exalytics In-Memory Machine и Oracle Big Data Appliance свидетельствуют о несомненных преимуществах этих оптимизированных программно-аппаратных комплексов перед традиционными, что открывает широкие перспективы для продвижения наших решений с использованием технологий Больших данных среди компаний финансового, телекоммуникационного сектора, в ритейле и других индустриях»

Алексей Голосов, президент ФОРС

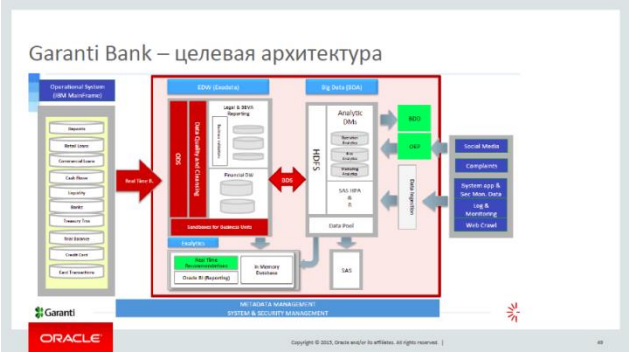
Примеры проектов

Oracle Big Data: референсные заказчики

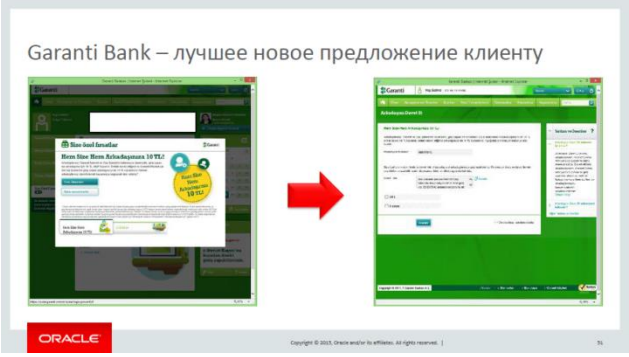
Garanti Bank

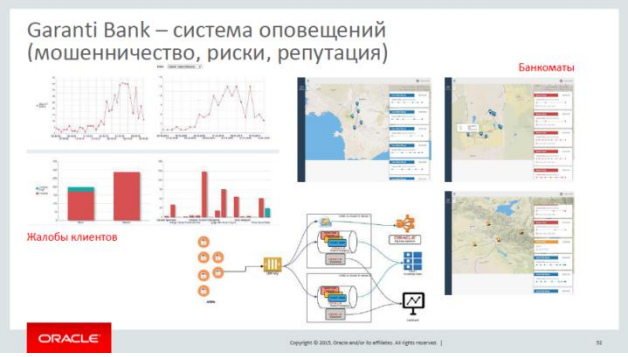


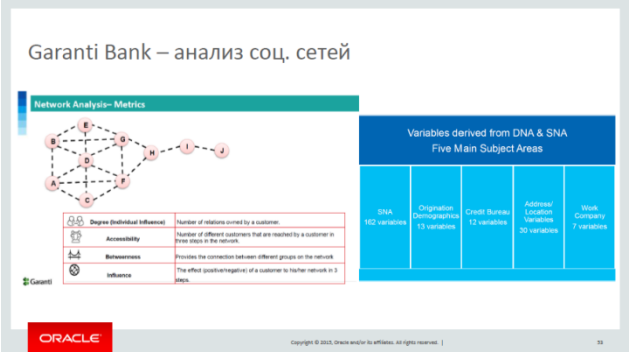

Garanti Bank – текущая архитектура
ORACLE

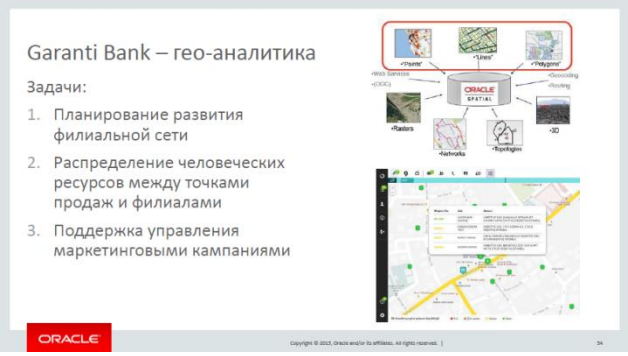

Garanti Bank – целевая архитектура
ORACLE


Garanti Bank – источники данных
Data Pool
Structured Data
Semi Structured
Unstructured
Credit Card Behaviour
Customer Financials
Customer Behaviour
Credit Bureau
Smart Application location
System and Application Logs (Netwitness , Arcsight, RUM) logs
Channel logs (ATM, Internet, IVR, CC)
Click Stream Data
Customer Complaints
Social Media data
Speech – to – text
ORACLE


Garanti Bank – лучшее новое предложение клиенту
ORACLE


Garanti Bank – система оповещений (мошенничество, риски, репутация)
Банкоматы
Жалобы клиентов
ORACLE


Garanti Bank – анализ соц. сетей
Network Analysis– Metrics
Degree (Individual Influence)
Accessibility
Betweenness
Influence
Variables derived from DNA & SNA
Five Main Subject Areas
SNA
162 variables
Origination Demographics
13 variables
Credit Bureau
12 variables
Address/ Location Variables
30 variables
Work Company
7 variables
ORACLE


Garanti Bank – гео-аналитика
Задачи:
1. Планирование развития филиальной сети
2. Распределение человеческих ресурсов между точками продаж и филиалами
3. Поддержка управления маркетинговыми кампаниями
ORACLE


ORACLE

DEEP LEARNING IN BIG MEDICAL IMAGE DATA

V. KOVALEV, PhD
Head of the Laboratory of Biomedical Images Analysis

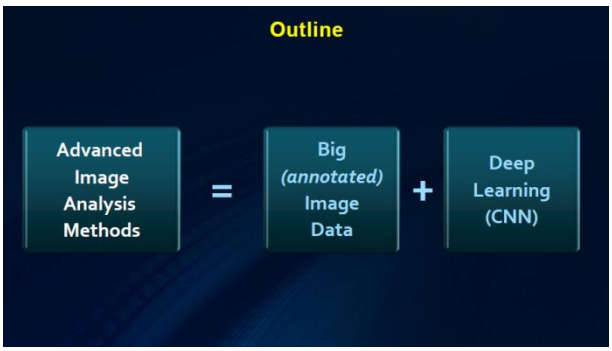
V. LIAUCHUK
Research Assistant of the Laboratory of Biomedical Images Analysis

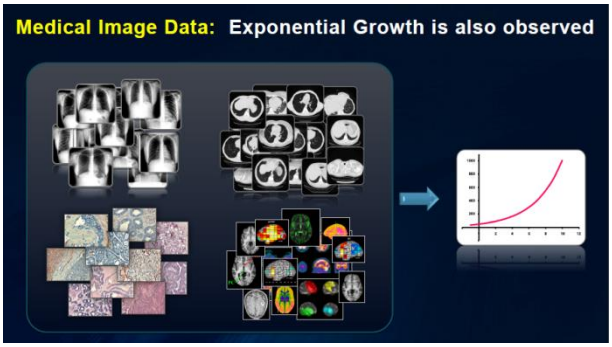
A. KALINOVSKY
Research Officer of the Laboratory of Biomedical Images Analysis

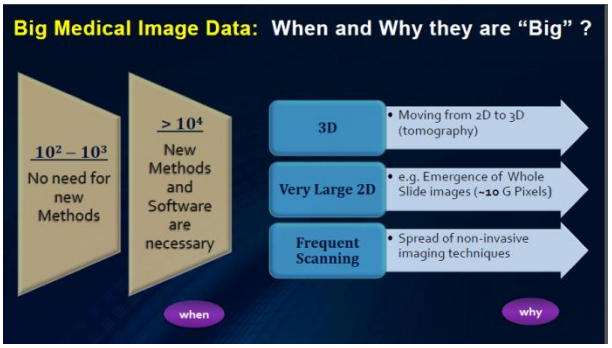
E. SNEZHKO
Senior research associate of laboratory of Mathematical cybernetics

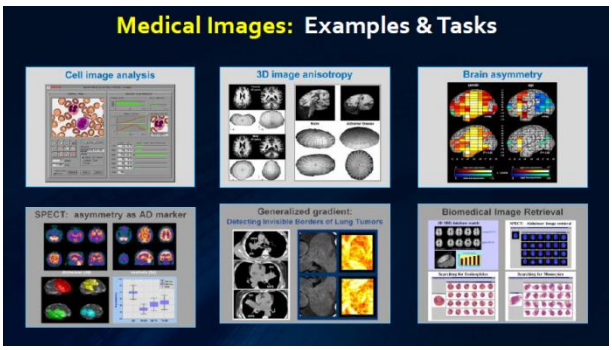
A. TUZIKOV, DSc.
General Director United Institute of Informatics Problems, National Academy of Sciences of Belarus, doctor of physical and mathematical sciences, professor, Corresponding member of the National Academy of Sciences of Belarus on "Informatics in Medicine and Biology"

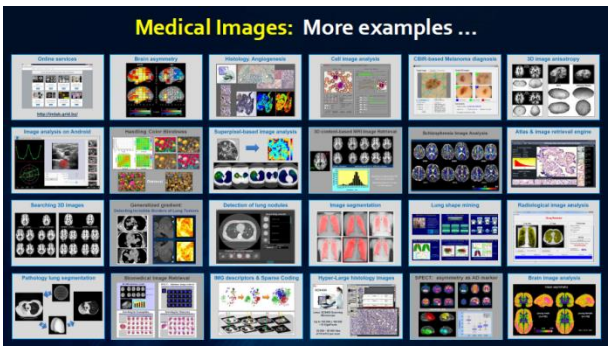
Biomedical Image Analysis Department, United Institute of Informatics Problems, National Academy of Sciences of Belarus, Republic of Belarus
E-mail: vassili.kovalev@gmail.com

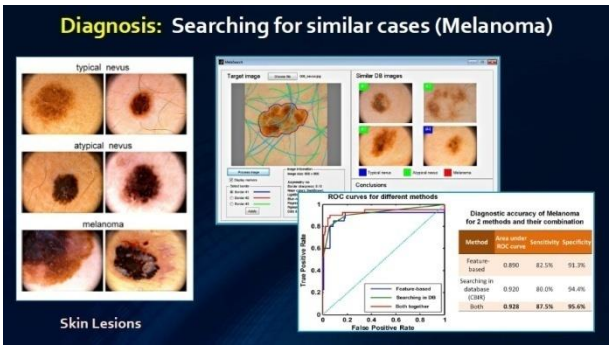


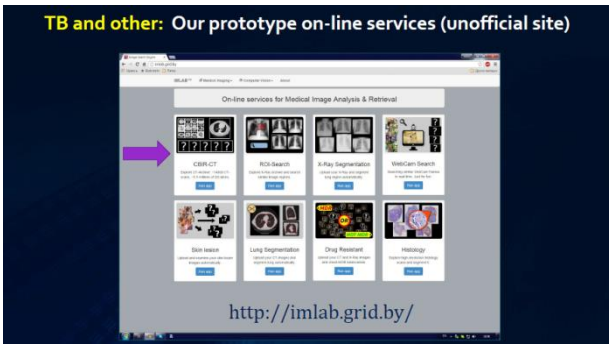

Medical Image Data: Exponential Growth is also observed

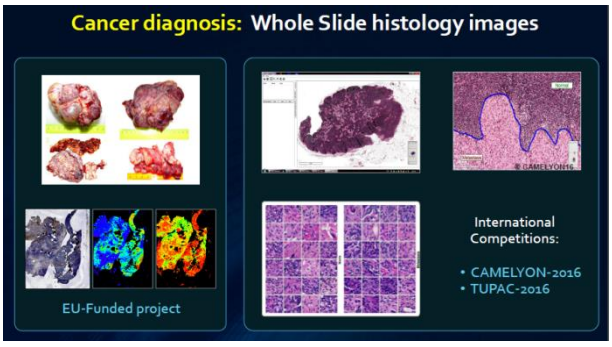

Big Medical Image Data: When and Why they are "Big" ?
$10^2 - 10^3$
No need for new Methods
$> 10^4$
New Methods and Software are necessary
3D
Moving from 2D to 3D (tomography)
Very Large 2D
e.g. Emergence of Whole Slide images (~10 G Pixels)
Frequent Scanning
Spread of non-invasive imaging techniques
when
why

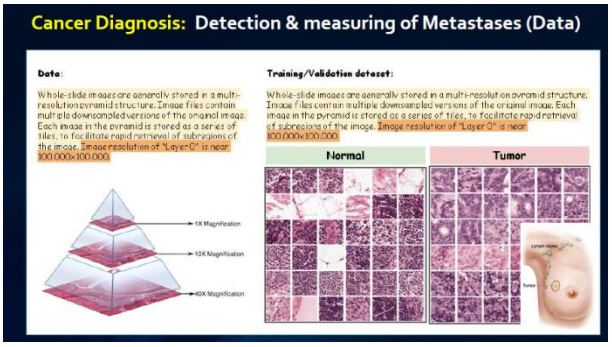

Medical Images: Examples & Tasks
Cell image analysis
3D image anisotropy
Brain asymmetry
SPECT: asymmetry as AD marker
Generalized gradient:
Biomedical Image Retrieval

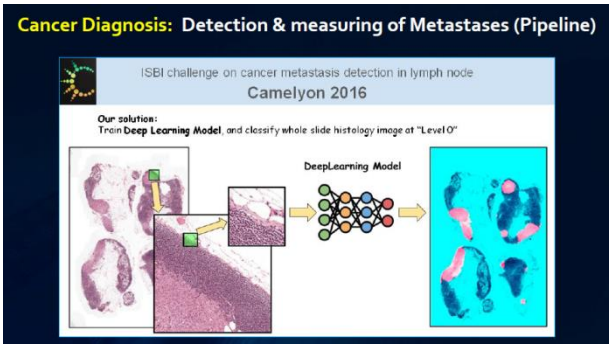

Medical Images: More examples ...

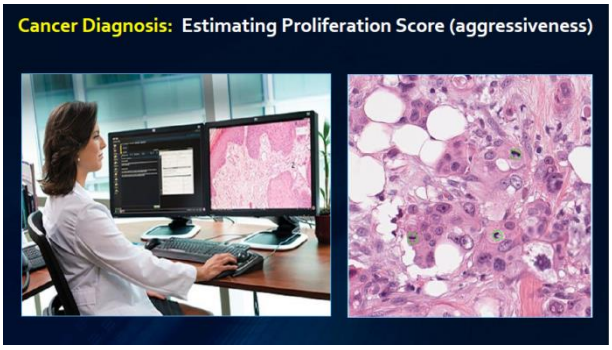

Diagnosis: Searching for similar cases (Melanoma)
typical nevus
atypical nevus
melanoma
Skin Lesions
ROC curves for different methods
Diagnostic accuracy of Melanoma for 2 methods and their combination

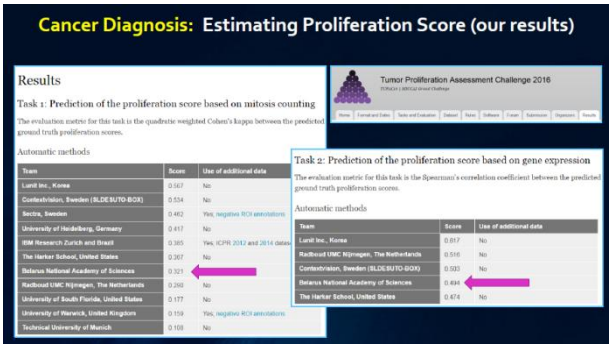

TB and other: Our prototype on-line services (unofficial site)
On-line services for Medical Image Analysis & Retrieval
http://imlab.grid.by/

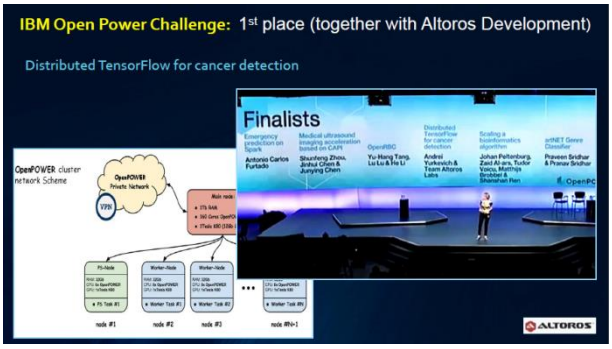

Cancer diagnosis: Whole Slide histology images
EU-Funded project
International Competitions:
• CAMELYON-2016
• TUPAC-2016

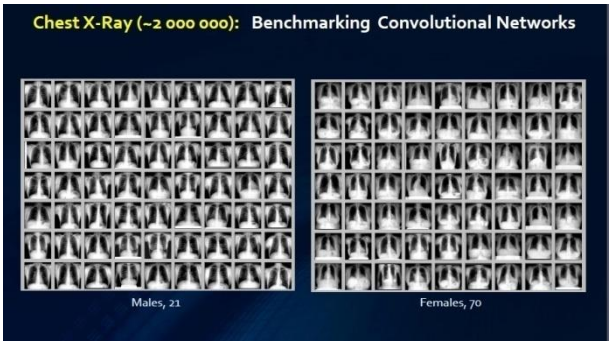

Cancer Diagnosis: Detection & measuring of Metastases (Data)
Data:
Training/Validation dataset:
Normal
Tumor

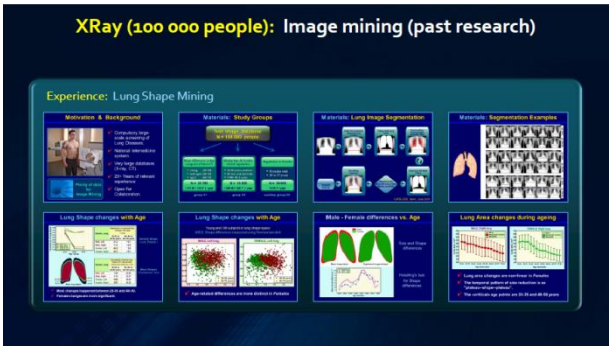

Cancer Diagnosis: Detection & measuring of Metastases (Pipeline)
ISBI challenge on cancer metastasis detection in lymph node
Camelyon 2016
Our solution:
Train Deep Learning Model, and classify whole slide histology image at "Level 0"
DeepLearning Model

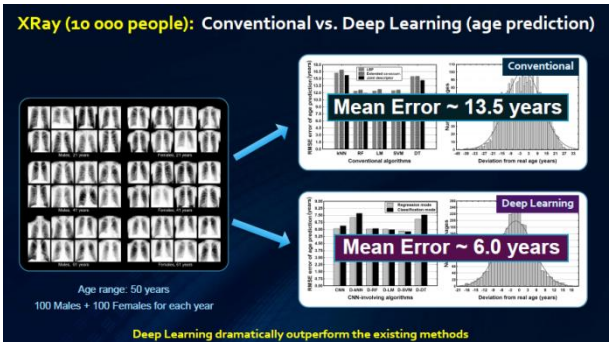

Cancer Diagnosis: Estimating Proliferation Score (aggressiveness)

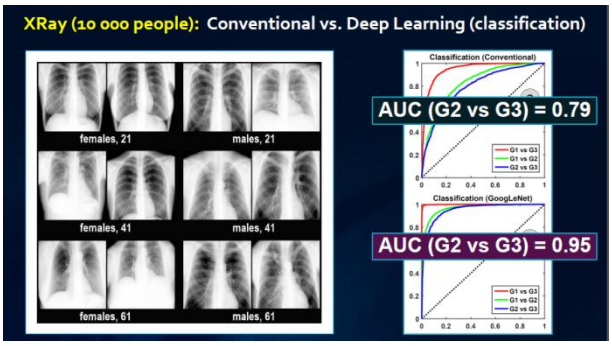

Cancer Diagnosis: Estimating Proliferation Score (our results)
Results
Task 1: Prediction of the proliferation score based on mitosis counting
Tumor Proliferation Assessment Challenge 2016
Task 2: Prediction of the proliferation score based on gene expression
Automatic methods

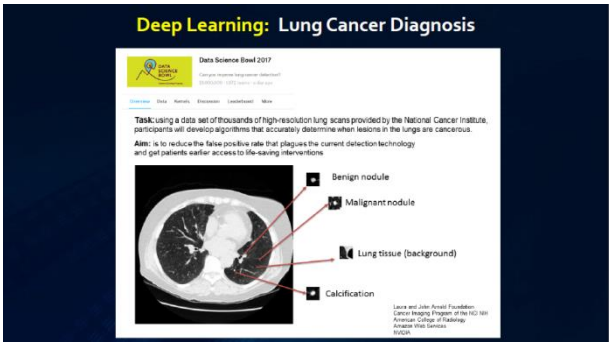

IBM Open Power Challenge: 1st place (together with Altoros Development)
Distributed TensorFlow for cancer detection
Finalists
OpenPOWER cluster network Scheme
ALTOROS

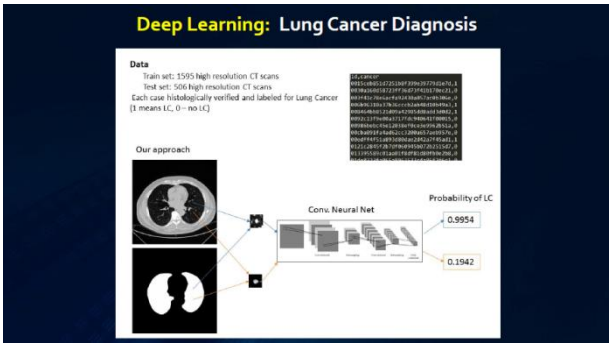

Chest X-Ray (~2 000 000): Benchmarking Convolutional Networks
Males, 21
Females, 70

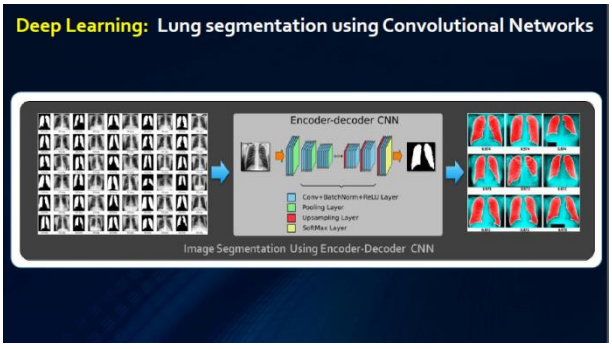

XRay (100 000 people): Image mining (past research)
Experience: Lung Shape Mining

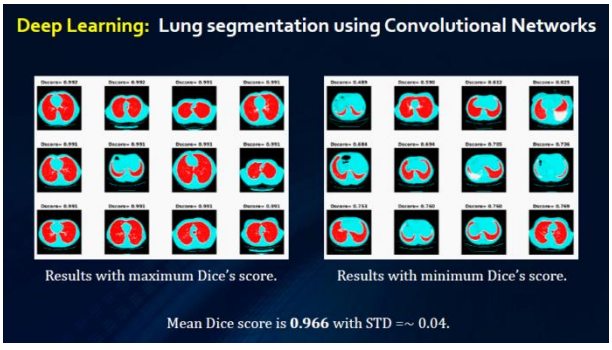

XRay (10 000 people): Conventional vs. Deep Learning (age prediction)
Conventional
Mean Error ~ 13.5 years
Deep Learning
Mean Error ~ 6.0 years
Age range: 50 years
100 Males + 100 Females for each year
Deep Learning dramatically outperform the existing methods

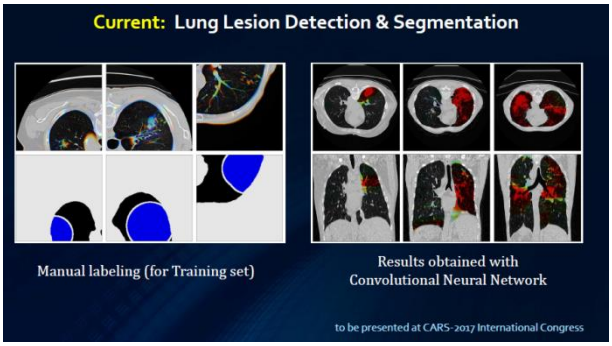

XRay (10 000 people): Conventional vs. Deep Learning (classification)
females, 21
males, 21
females, 41
males, 41
females, 61
males, 61
Classification (Conventional)
AUC (G2 vs G3) = 0.79
Classification (GoogLeNet)
AUC (G2 vs G3) = 0.95

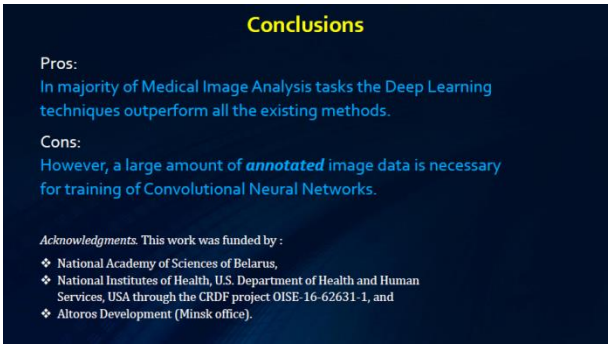

Deep Learning: Lung Cancer Diagnosis
Data Science Bowl 2017
Benign nodule
Malignant nodule
Lung tissue (background)
Calcification


Deep Learning: Lung Cancer Diagnosis
Data
Train set: 1595 high resolution CT scans
Test set: 506 high resolution CT scans
Each case histologically verified and labeled for Lung Cancer
(1 means LC, 0 – no LC)
Our approach
Conv. Neural Net
Probability of LC
0.9954
0.1942


Deep Learning: Lung segmentation using Convolutional Networks
Encoder-decoder CNN
Conv+BatchNorm+ReLU Layer
Pooling Layer
Upsampling Layer
SoftMax Layer
Image Segmentation Using Encoder-Decoder CNN


Deep Learning: Lung segmentation using Convolutional Networks
Results with maximum Dice's score.
Results with minimum Dice's score.
Mean Dice score is 0.966 with STD =~ 0.04.


Current: Lung Lesion Detection & Segmentation
Manual labeling (for Training set)
Results obtained with Convolutional Neural Network
to be presented at CARS-2017 International Congress


Conclusions
Pros:
In majority of Medical Image Analysis tasks the Deep Learning techniques outperform all the existing methods.
Cons:
However, a large amount of annotated image data is necessary for training of Convolutional Neural Networks.
Acknowledgments. This work was funded by :
National Academy of Sciences of Belarus,
National Institutes of Health, U.S. Department of Health and Human Services, USA through the CRDF project OISE-16-62631-1, and
Altoros Development (Minsk office).

APPLICATION OF TIME SERIES TO PERFORMANCE ASSURANCE OF BIG DATA ENVIRONMENT

Y. BALASANOV , PhD,
University of Chicago

B. ZIBITSKER, PhD
President and CEO
BEZNext, Emeritus professor of BSUIR

T. BAKANAS
University of Chicago Graham School of Continuing Liberal and Professional Studies

E. HAMMOND
University of Chicago Graham School of Continuing Liberal and Professional Studies

M. ISLAS-MARTINEZ
University of Chicago Graham School of Continuing Liberal and Professional Studies

CEO BEZNext, Chicago, USA
E-mail: bzibitsker@beznext.com

Abstract. The selection of the Big Data algorithms, YARN rules and infrastructure can affect accuracy, performance and scalability of Big Data Applications.

We will present a methodology and algorithms for proactive performance management. Every hour collected measurement data are aggregated into workloads representing each lines of business. Each workload has three profiles, including 1) performance (response time and throughput), 2) resource utilization and 3) data usage profiles. Profiles represent Workloads' Time series. This information is used as input for exploratory analysis techniques specific to time series data. The data are transformed into stationary Time Series and an analysis to select the best time series model (ARMA, VARMA) is conducted. Historical data are used to identify past exceedances which are utilized as predictors or outcome variables to build a classification model.

We will review short term prediction, seasonal peaks identification, diagnostic and root cause analysis Performance Assurance algorithms enabling proactive performance management of Big Data Application.

Key Words: Missing data, time series data, anomaly detection, performance prediction, big data, performance assurance.

*Introduction.*Big Data applications implementation success depends on design, implementation and management decisions.

Selection of Machine Learning (ML) algorithms and libraries, software timing parameters and rules assigning priorities and concurrency limitations. for software subsystems like YARN, Kafka, Spark, Storm and selection of the architecture and hardware configuration also can affect performance, scalability and cost.

Difficulty of managing complex multi-tier, distributed, virtualized, parallel processing environment creates an uncertainty and high risk of performance surprises for Big Data applications.

Failure to meeting Business Service Level Goals (SLG) can affect ability of business to make real time decisions. Therefore, our goal is to implement performance prediction technology justifying proactive measures necessary to meeting SLGs in growing and constantly changing environment.

Predictive models are based on performance measurement data continuously generated by each of Big Data subsystems. It includes response time and throughput, resource utilization, (CPU, memory, storage and network) by Systems, Users and Applications and Data usage, including read/write ratio, level of parallelism during accessing data. We aggregate measurement data into hourly profiles by business workloads / line of business.

Each workload use different set of applications. The number of users and volume of data are growing. The number of new applications is increasing as well. It increases the contention for resources and affects performance.

We aggregated performance measurement data into 44 business workloads, characterized by performance, resource utilization and data usage as a time series. All workloads compete for the same resources and affect each other performance, so they should be tested for the degree of cointegration

to determine the type of time series analysis that can be conducted. Different applications are active during different time of the day. Hardware and software configuration periodically changes as a result of tuning or hardware outages. Many workloads have a random arrival rate and service time [Buzen].

Gaps in measurement data and randomness of the performance of workloads characterization affect the selection and use of the modeling algorithms.

Queueing network models for example, assume the Poisson distribution of the independent variables. Calibration of the multi-tier distributed, virtualized parallel processing environment supporting mix workloads have many challenges [z and S and L]. ML algorithms which we implemented use a lot of resources and do not scale well.

In this paper we will illustrate how time series algorithms can be used for performance prediction, determining seasonality and performing the diagnostics and root cause analysis to justify Performance Assurance decisions and implement dynamic performance management. We will review problems of data preparation, fitting and applying time series and categorical models. We will present a methodology of selecting the segment from an incomplete time series that maximizes the number of original data points while minimizing the amount of data that need be imputed, review a point forecasting model for each performance indicator for each workload to detect and predict extreme positive anomalies in key performance variables

Data Collection, Aggregation and Preparation. BEZNext agents continuously collect measurement data from Linux, YARN, Kafka, Storm, Spark, Cassandra and other Big Data subsystems [BZ]. Measurement data are aggregated by workloads in hourly performance, resource utilization and data usage profiles to represent the aggregated activity of users and applications supporting each line of business. [BZ]

Data set used during this study contains 44 workloads with 15 variables each. List of Variables includes: Date, Hour, Workload Name, Number of Parallel Jobs, Number of Parallel Requests, Number of Network Messages, Total Number of Execution, Total Arrival Rate, Average Response Time, Total CPU Time, Total # IO, Total Delay Time, Total CPU Utilization, Concurrency Limit, CPU Utilization Limit.

Any changes in number of users, changes in volume of data, implementations of new applications or modifications of existing applications and corresponding changes in performance and usage of resources are reflected in the hourly workload profiles. The Average Response Time and Total Number of Executions per hour are the Key Performance Indicators (KPIs), which are used in the time series and categorical models. All other variables are used as predictors for the KPI.

The methodology presented in this paper focus on data imputation, time series forecasting, seasonal peaks and anomalies determinations, performance problems and their root causes prediction and development proactive performance management recommendations.

Imputation. Time series data have a significant number of time intervals without observations.

In a 2010 paper in *the American Journal of Political Science*, two researchers, Honacker and King, apply a particularly apt analogy to the problem. If a data set is a cheese, most real data sets are Swiss, filled with holes. With some data sets, one could simply drop the missing data pieces. When dealing with time series data, however, the missing chunks cannot be dropped as both auto-regressive and moving average predictive techniques depend upon consecutive values.

Since for time series data melting the cheese down and reforming with no holes doesn't work, the other option is to fill the holes. Filling in long chunks of data, however, runs the risk of either artificially reinforcing or removing any seasonality in the time series. As Honacker and King put it, "if you fill the holes in the cheese with peanut butter, you should not pretend to have more cheese!" (Honacker and King, 2010).

Honacker and King suggest a method of data filling that is based upon building an imputation model and combining the imputed values with expected values from historical analysis of similar data sets (Honacker and King, 2010). Their work, while robust, does not cleanly apply to the problem of

Big Data systems performance forecasting, where some of the workloads might be not active for some time and some of the nodes might having problems and not available during period of time. Imputation over such a long period loses its validity.

We applied algorithm that selects a time frame based on maximizing length while minimizing hours filled is selected as the best method of imputation for this data. In terms of the cheese analogy mentioned in the Background section, it identifies as large a portion of the Swiss as it can where all the holes all remain under a certain size. The idea is that small holes can be reliably filled with a simple algorithm without biasing the data. Using this methodology can produce unbiased data sets capable of supporting further analysis.

The algorithm developed has two parameters, neighborhood size and minimum size. The neighborhood size is the number of values before and after a missing value that will be used for imputation. The minimum size is the minimum length of the selected segment. There are cases where a Workload has multiple segments that meet these requirements. In those cases we select the most recent segment.

The proposed methodology is described below:

- Select a time series and identify all missing values within
  - For our analysis, each workload is examined over the length of one year.
- For each missing value extract a neighborhood of adjacent values.
  - Our analysis used a total neighborhood size of ten, which includes the five values preceding the missing point and five after.
- Each missing value is evaluated.
  - If there is at least one true value on both sides of the point within the neighborhood, then missing value is imputed with average of all the existing values within the neighborhood.
  - If there are no existing values on both sides of the point, then the point remains missing.
  - If a value cannot be imputed, then the segment is ended and a new segment starts at the next non-missing value.
- The best subset is selected.
  - The short-term analysis uses the most recent segment.
  - The anomaly prediction analysis uses the largest segment.

This subset selection methodology allows for the consistent treatment and cleansing of time series data across many distinct time series with differing degrees of missing data.

In the graphs below we see how the imputation algorithm selects both the ‘largest’ and the ‘best’ segment for the Workload DBC. For DBC the ‘Largest’ segment occurred from 2015-05-03 through 2015-09-18. The imputation algorithm selected a different, more recent, time frame for the ‘Best’ segment. Both perform as they are supposed to, a larger set of data is selected to feed the anomaly prediction from ‘Largest’ and the most recent set of data is selected to feed the time series model from ‘Best.’

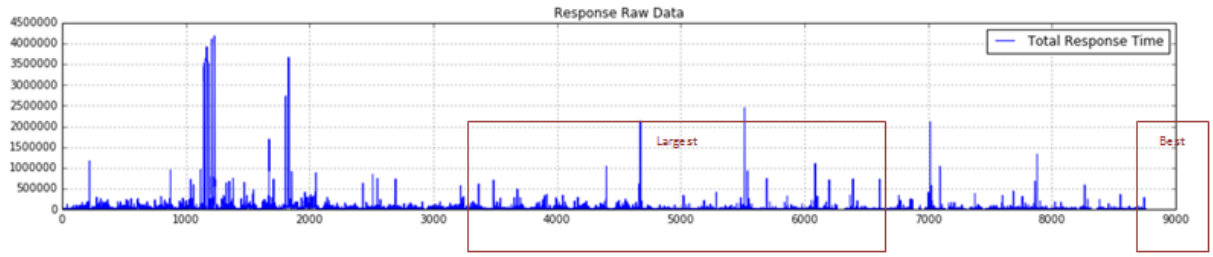


Fig. 1. DBC Workload Total Response Time Full Year

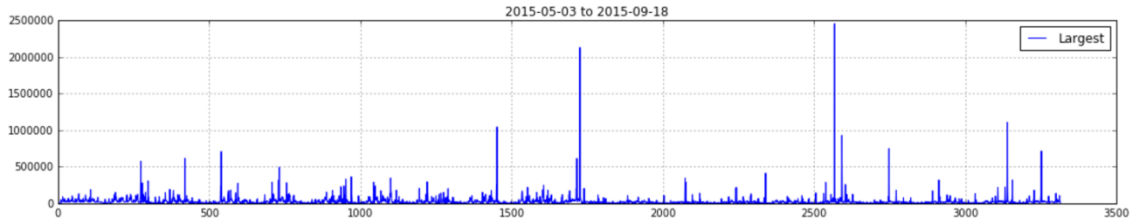


Fig. 2. DBC Workload Total Response Time Largest Segment

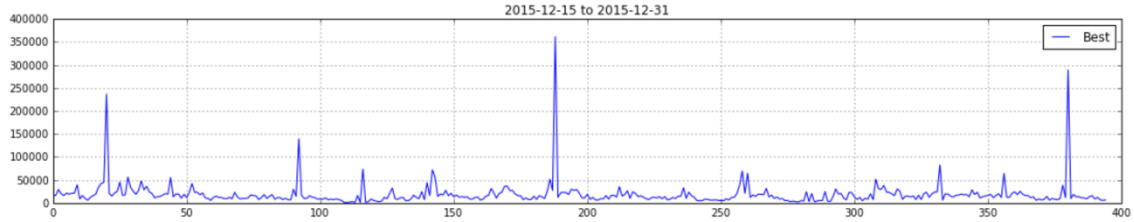


Fig. 3. DBC Workload Total Response Time Best Segment

The Workload Best and Longest segments have return a variety of lengths across the 44 Workloads. In many cases a segment from a Workload is selected as both the Best and the Longest. As can be seen from the histograms below the 'Best' segments trend shorter than the 'Longest' segments. This is due to the requirement that the 'Best' segment be the more recent available.

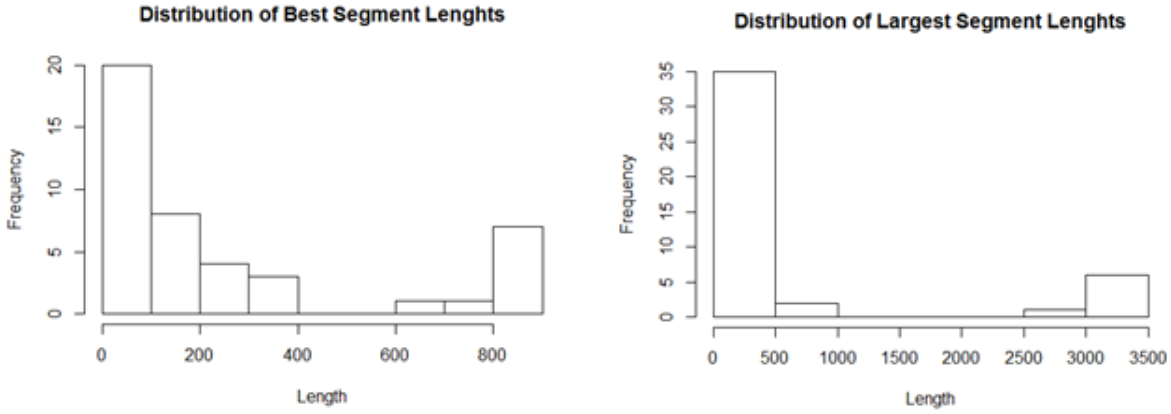


Fig. 4. DBC Workload distribution of best and largest segment lengths

The imputation methodology is reliable and flexible. From 44 Workloads with varying data quality it selected time periods capable of supporting two types of analysis. We used methods based on linear Time Series models which do not require very long samples.

Time Series Forecasting. Big Data systems produce multiple time series system status variables that are likely to be related. Aboagye-Sarfo, Mai, Snfilippo, Preen, Stewart and Fatovich (2016) present VARMA forecast model as a technique designed for modeling multiple time series simultaneously. Aboagye-Sarfo et all use VARMA for predicting emergency department demand in Western Australia. They compare results that use benchmark univariate autoregressive moving average (ARMA) and Winters' techniques. Their research demonstrates that a VARMA model provides a more accurate forecast than the normal or standard univariate ARMA and Winters' methods.

We followed a similar methodology. After assessing stationarity for each variable in each workload a VARMA and ARMA model is fit to each of the KPI variables. We expect the VARMA model

to be the strongest fit, but by running both time series techniques we are able to customize the results for each variable workload combination..

Example of Time Series Forecasting for DBC workload

Overall, ARMA models had the best forecasting performance with all three models out-performing the naïve $Yt - 1$ forecasting model. Average Response Time had an MASE value over one which indicates that using $t - 1$ for a prediction would have been more accurate than the VARMA model.

Table 1. DBC 1-Hour Forecast MASE

Variable	ARMA	3-VARMA
Average Response Time	0.9439	1.240
Total IO	0.8184	0.835
Total CPU	0.9268	0.896

Extreme outliers are a major obstacle in our forecasts, with the VARMA models proving to be particularly sensitive. The figures below show that the VARMA forecasts an extreme outlier around hour 85, where none occurs in the original data.

Stationarity. The first step in building a time series forecasting model is testing for stationarity in the data. If the time series is not stationary, then it must be transformed into a stationary state otherwise the model will not be stable enough to generate accurate forecasts. Testing for stationarity can be done quantitatively using statistical tests such as the Augmented Dicky-Fuller (ADF) and the Kwiatkowski–Phillips–Schmidt–Shin (KPSS) tests. These tests evaluate a time series for the presence of a unit root or trend-stationarity. Stationarity testing can also be done be qualitatively by examining autocorrelation plots for exponential decay.

Numerous techniques (de-trending, differencing, seasonal adjustment, log transformation, smoothing) can be used to transform non-stationary data into stationary data. This analysis uses the differencing and seasonal adjustment techniques to achieve stationarity in the data.

Overall, the stationarity tests indicate that our differencing selection methodology produced accurate results. Before each model was constructed ADF and KPSS tests verified that the data used to train the series models was stationary. No variables were found to be non-stationary after differencing.

Differencing. A stationarity transformation algorithm was developed for this analysis. The algorithm is a brute force approach that differences each variable with a set of predefined differencing functions. Then, metrics are calculated from the differenced data sets and ranked. The differencing functions were selected based on our knowledge of the problem domain and segment lengths identified in the imputation stage. The differencing techniques evaluated for each variable are:

- No difference.
- First-order difference.
- Second-order difference.
- Daily difference.
- Weekly difference.
- Monthly difference.
- First-order + daily difference.
- First-order + weekly difference.
- First-order + monthly difference.

Each differencing function is applied to each variable of each Workload and a set of ranking parameters are calculated. These parameters are used to select the optimal differencing function. The ranking parameters are

- Mean absolute scaled error of ARMA in-sample predictions after integration (lower is

better).

- Root mean squared error of ARMA in-sample predictions after integration (lower is better).
- Count of autocorrelation function points with the confidence interval (higher is better).
- Count of partial autocorrelation function points with the confidence interval (higher is better).
- Absolute average of autocorrelation function points (lower is better).
- Absolute average of partial autocorrelation function points (lower is better).
- The differencing function is automatically rejected if one of the Augmented Dicky Fuller, Kwiatkowski–Phillips–Schmidt–Shin and Phillips–Perron tests indicate nonstationary.

This brute force process allows for many independent time series with different types of stationarity and seasonality to be accounted for in an automated fashion. Each differencing function also has a corresponding integration function that is used to transform the forecasts generated by the ARMA/VARMA models. Refer to Appendix 2 for example Python code of the differencing functions used in this analysis.

Five unique differencing functions were selected. However, the most commonly selected approach was to use no differencing. The most common differencing function was first-order differencing. Thirteen variables had seasonality detected by our algorithm.

Table 2. Differencing Functions Selected

Function	Count
no_difference	52
first_order_difference	41
weekly_difference	7
daily_difference	5
second_order_difference	3
first_order_weekly_seasonal_difference	1

For instance, the differencing algorithm determined that the three variables in the DBC Workload were already stationary and selected the “No Difference” differencing function. In contrast, the Total CPU variable of the SALES Workload was non-stationary. We will illustrate the differencing selection algorithm by using the SALES Workload.

The autocorrelation function of the SALES Workload indicates non-stationarity and possible seasonality.

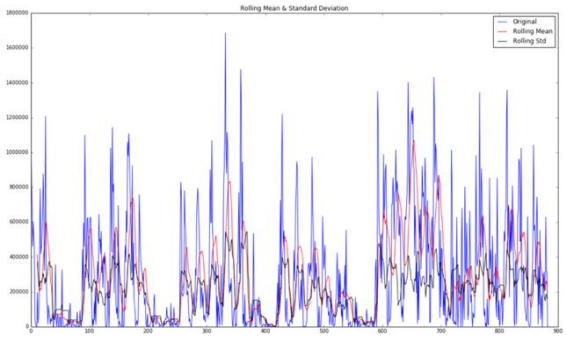


Fig. 5. SALES Workload Total CPU raw data

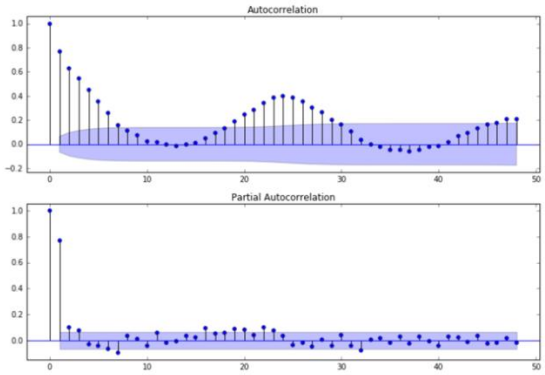


Fig. 6. SALES Workload Total CPU ACF and PACF plots

The periodogram is extremely noisy, with no clear indication of seasonality.

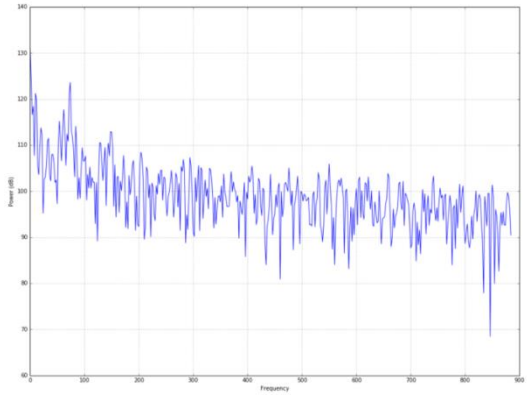


Fig. 7. SALES Workload Total Periodogram plot

Again, there is a disagreement between the periodogram and the autocorrelation plots. Auto-correlations indicates that there is daily seasonality, while the periodogram does not. The differencing selection algorithm selected “First-Order Difference” as the differencing function.

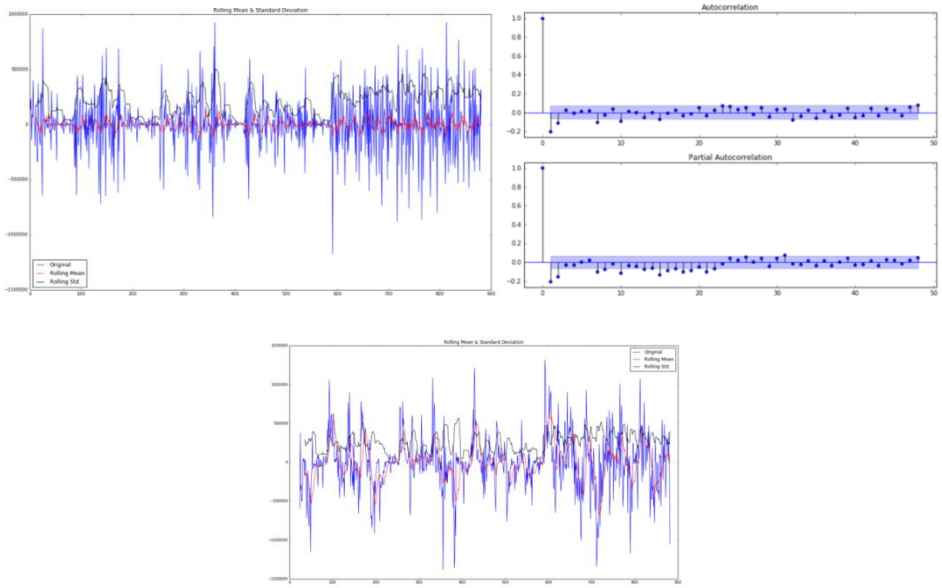


Fig.8. SALES Workload Total CPU First-Order Difference

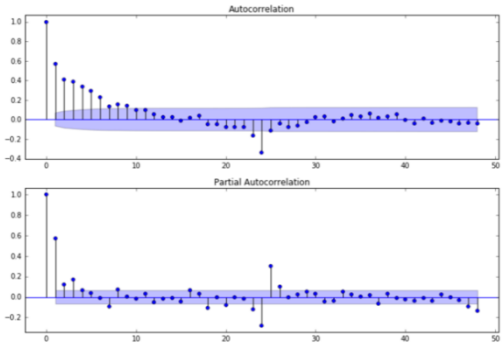


Fig. 9. SALES Workload Total CPU Daily Difference

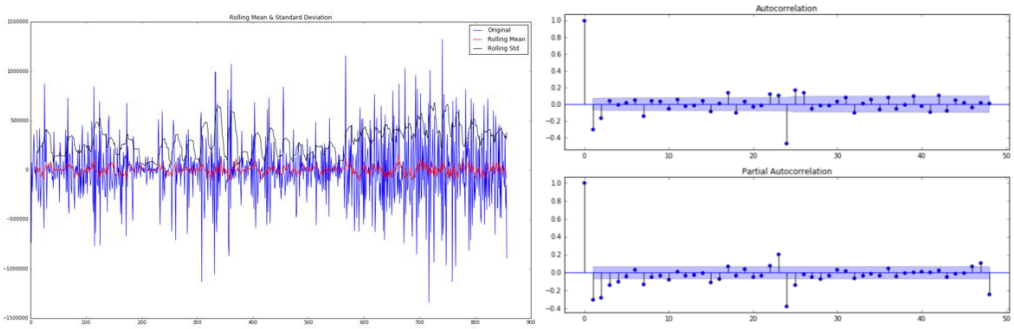


Fig. 10. SALES Workload Total CPU First-Order + Daily Difference

Three autocorrelation plots above show that the differencing selection algorithm correctly identified the best differencing technique. The “First-Order” differencing function has the best autocorrelation plot of the three.

Cointegration. We hypothesize that the Workload data from BEZNext contains variables that are dependent on each other. This lends itself to VARMA time series analysis. A requirement of a VARMA model, however, is that data be cointegrated.

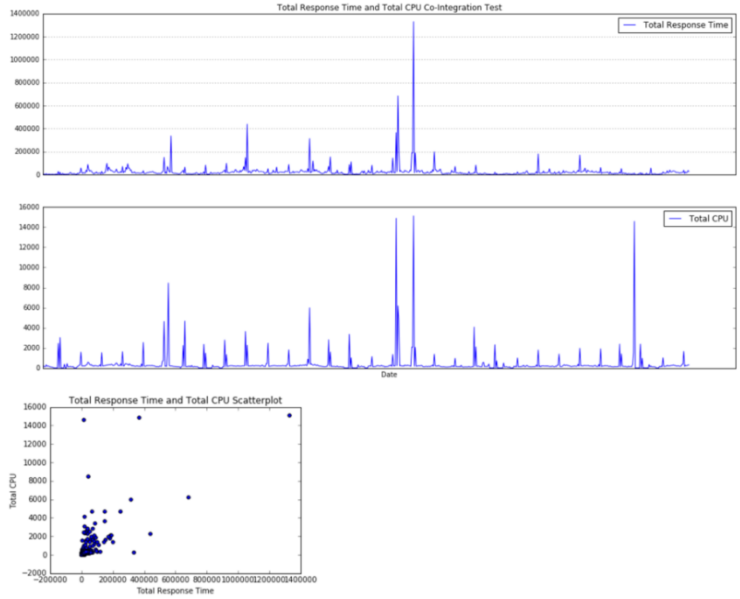


Fig. 11. DBC Workload response time vs CPU Cointegration results

Our approach uses the Augmented Dickey-Fuller and the Johansen tests to determine if any variables are cointegrated. These are standard tests that are applicable to a wide range of time series data.

We find that the Key Performance Indicators mostly are cointegrated with each other. In some Workloads, the Average Response Time Variable is not cointegrated and is excluded from the VARMA time series model.

The ADF and Johansen tests both show cointegration between the CPU and IO variables. We run these tests for each Key Performance Indicator before building the VARMA model for each Workload. Any analysis of Workload data from Big Data systems should undergo testing to determine if its descriptor variables display cointegration.

Model Construction. ARMA and 3-variable VARMA forecasting models are constructed for each variable, Both models follow a similar workflow:

1 If VARMA, test for cointegration.
2 If necessary, difference data.
3 Train models for all *p, q* combinations between 0, 0 and 5, 5.
4 Select best *p, q* values using AIC as the evaluate metric.
5 Generate forecasts.
6 If using differenced data, integrate previous value(s).

The univariate ARMA approach is the most straight-forward model and directly follows the above workflow. One model is constructed for each of the Key Performance Indicators in each Workload. This approach has the fastest training and cross validation times by a large margin.

VARMA forecasts are generated for variables without cointegration. A single VARMA model is generated for each Workload.This approach has slower training and evaluation times compared to ARMA.

This approach constructs three VARMA models per Workload, one for each Key Performance Indicator, with each model containing the fields cointegrated with the target Key Performance Indicator. This approach seems to the most through and our initial hypothesis was that this technique would be the most accurate method. However, this approach proved to be very cumbersome, with extremely long training and evaluation times. The large number of variable combinations often resulted in models with invalid states. Many models did not converge or had errors when fitting the model to the data.

Refer to Appendices 3 through 5 for Python code samples showing the model construction and cointegration testing.

Model Evaluation. Cross validation evaluates the forecasting accuracy of the ARMA and VARMA models. A rolling-origin forecast cross validation procedure follows these steps:

1 Create an 80-20 train test split within each Workload time series.
2 Train the model on the train data set.
3 Forecast $T + 1$ value.
4 Roll the origin forward in the train data set by adding the Salesl $T + 1$ value.
5 Retrain the model.
6 Forecast $T + 2$ value.
7 Continue until all values in the test dataset have been forecasted

We find the VARMA usually outperforms the ARMA model in the rolling forecast origin cross validation. However, we also want to see how these forecasts perform with multiple forecasts instead of just forecasting a single hour. We look at 3 to 6 hour forecast periods. The forecast accuracy decreases as it predicts further to the future.

We also noticed that in some cases accuracy increases proportionally to the forecasting intervals. In these cases, the models revert to predicting a constant value around 4 hours in the future. This lead to accurate forecasts in T + 4 or later if we had properly selected the differencing function and there were not a significant number of anomalies in the data.

Seasonality. Generally, qualitative approaches are used for identifying seasonality in time series data. Patterns indicating seasonality can often be spotted in the power spectral density or autocorrelation plots after careful review. However, these approaches can result in inconsistent interpretations and tend to be very time-consuming.

A periodogram plot of a time series' power spectral density (PSD) is one of the most common ways of identifying seasonality. When focused on a single time series, techniques can be applied to draw a signal out of noisy data (Fernandez et all, 2016). However, proper interpretation can be extremely challenging. The plot below shows the Total IO variable from the DBC Workload. At approximately daily intervals there are spikes in the time series which indicates that there may be some seasonality in this data.

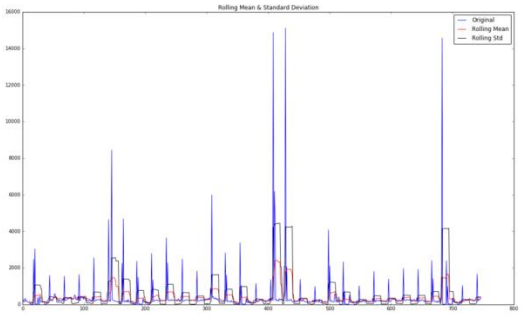


Fig. 12. DBC Workload Total CPU raw data

The periodogram below indicates that there might be a daily seasonal pattern as well. The left plot is the standard power spectral density while the plot on the right is the transformed PSD plot with hours on the X-axis instead of frequency. There is a large spike at hour 24 in the transformed plot indicating a daily seasonal pattern.

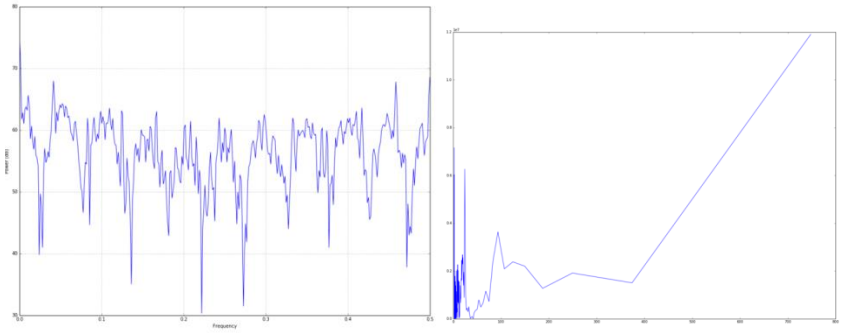


Fig. 13. DBC Workload Total CPU Periodogram

An autocorrelation plot can also be used to identify seasonality. The autocorrelation and partial autocorrelation plots are shown below and clearly indicate that there is no seasonal pattern. In fact, the data already appears to be stationary with very little autocorrelation.

The above example shows how difficult it can be to perform these qualitative interpretations. The raw data seems to contain some form of seasonality. The periodogram while very noisy shows some element seasonality may be present. However, the autocorrelation plot is a stark contrast. It shows no indication of seasonality at all.

The other challenge in identifying seasonality is the sheer number of Workloads in our data set. Evaluating only the Key Performance Indicators in each of the 44 Workloads requires 132 separate analyses and a full evaluation of all variables requires 528 analyses.

Consistent interpretation in a timely manner requires an automated approach. Differencing can

be used to both transform a time series into a stationary form and to capture any seasonal patterns if they exist. The key is to select a differencing function that addresses both conditions.

Identification of the workload's seasonality enables organization of the proactive planning of changing workloads' management rules, including priorities, concurrency and resource allocation.

Diagnostic and Root Cause Analysis. Many businesses are making business decisions in real time. It creates a pressure to develop real time Big Data applications capable to extract necessary attributes from the stream of data supported by Kafka or Flume and apply ML algorithms processing data by Spark or Storm in memory. It is critical to be able to detect anomalies and predict performance problems and their root causes in order to proactively and dynamically make necessary changes to continuously meet SLGs for each workload. Alexander Lavin and Subutai Ahmad ("Evaluating Real-time Anomaly Detection Algorithms – the Numenta Anomaly Benchmark", 2015) propose a benchmark anomaly detection methodology called Numenta Algorithm Benchmark (NAB). The technique applies a variety of anomaly detection and prediction algorithms and selects the combination it deems has the highest score. The score is determined by rewarding early detection while penalizing false results.

When deciding how to approach the anomaly detection and performance problems prediction, a real time streaming approach such has NAB was considered. However, because Workloads' variables are hourly aggregations the approximate real time and simpler techniques may be applied to achieve the same effect. Since models can be re-fit in the intervening hour between data point arrivals, past extreme positive anomaly detection and future problem prediction can be treated separately.

The anomaly detection approach followed in this paper focuses on detecting both global as well as local anomalies through applying the Generalized Extreme Studentized Deviate test. This technique can be used to identify one or more outliers in a data set. It holds the advantage that the number of outliers does not need be preset, but instead is determined within the algorithm.

In addition to NAB, a real time streaming model that combines Markov models and Bayesian classification methods can be used to predict anomalies. Experiments show that this approach efficiently predicts and diagnoses extreme positive anomalies with high accuracy (X. Gu and H. Wang, 2016). However, these models are highly complex, and due to the slightly less than real time nature of our data, such techniques need not be applied. This paper focuses on applying traditional classification models such as logistic and extremely randomized trees to the identified anomalies while using lagged variable data as the predictors.

For Big Data Applications, extreme positive anomalies on system status variables represent when resources are being stretched to capacity. The goal for this stage is to design a framework that enables automated prediction of these anomalies within Workload KPI variables and thus provide a warning to Big Data Application administrators of when changes to the configuration may be necessary.

Anomaly Detection. To train a model to predict anomalies, historical anomalies must first be identified. Two techniques from previous methodology sections are applied to generate a clean and continuous data set for anomaly detection. The imputation and best segment methodologies provide the data set for each Workload's anomaly detection. However, where the time series analysis used the most recent segment above a certain size, it was deemed more important for the anomaly detection to find the largest segment available to provide more anomalies against which to train.

This Generalized ESD (gESD) algorithm is then applied to the data set generated for each Workload and KPI variable. The gESD works best on data that approximates normal. The code package used automatically normalizes the Workload segment fed to it to meet this requirement. The generalized ESD algorithm assumes there can be up to n anomalies. The algorithm iterates by removing the point with highest G value (the point farthest away from the mean of the sample) calculated from the similarly iterated sample mean and standard deviation. The critical value changes with the number of points that are removed from the sample. The number of anomalies is the condition with the most outliers with a G above the critical value.

*Anomaly Prediction.*Now that the anomalies of each Key Performance Indicator are identified, the next step is to design an anomaly prediction model that will alert a company in time to change its system configurations. To do this, we run four probabilistic machine learning models for each workload. The dependent variables are a time series of binaries for each workload KPI variable: 1 if anomaly detected in that hour, 0 if not. The independent variables are time lagged values of all 15 variables. It is this lagging that necessitated the anomaly detection be done on a continuous set of data.

A lagged data frame is created for each Workload KPI variable (6 hours of lags for the DBC Workload, as the example on Table 2 outlines). This effectively increases the number of independent variables from 15 to 90. It also removes the data from its time series format and directly pairs each set of dependent and independent variables.

Table 3. Workload lagged data frame sample

	Anomaly	Average Response Time	Average Response Time_lag1	Average Response Time_lag2	Average Response Time_lag3	Average Response Time_lag4	Average Response Time_lag5	Average Response Time_lag6	Total Arr Rate	Total Arr Rate_lag1	...	Total Parallel Sessions_lag4	...
6	0	271.0	178.0	184.0	153.0	167.0	188.0	170.0	0.0	0.0	...	21.0	...
7	0	710.0	271.0	178.0	184.0	153.0	167.0	188.0	0.0	0.0	...	17.0	...
8	0	417.0	710.0	271.0	178.0	184.0	153.0	167.0	0.0	0.0	...	18.0	...
9	0	155.0	417.0	710.0	271.0	178.0	184.0	153.0	0.0	0.0	...	15.0	...
10	0	292.0	155.0	417.0	710.0	271.0	178.0	184.0	0.0	0.0	...	15.0	...

5 rows × 57 columns

A logistic regression model using all fields is implemented. Then, recursive feature elimination is conducted to trim the variables to only the most important. A new model is trained on the selected variables. It is tested on a 70/30 train/test data split.

The same sequence is conducted with an Extremely Randomized Trees classification model. The ERT model was selected over the standard random forest due to the faster training time and better classification accuracy (Geurts et all, 2006). The ERT is run using all 90 lag variables. Recursive feature selection is conducted on the results. Then the ERT is refit with just the selected variables and tested on a 70/30 train test split.

The logistic regression model does not always handle the skewed nature, few anomalies spread over many data points, of the datasets correctly. This results in high occurrences of false-negatives. The extremely randomized trees model handles the skewed dataset slightly better, reducing false-positives and false-negatives in many cases.

All the models are run for each Workload and the error rates are collected. The model with the lowest error rate is selected as best.

Anomaly Detection and Prediction Example. The plot below demonstrates an example of the gESD algorithm used with the DBC Workload data. The red points represent the anomalies that are detected. The anomalies are displayed against the undifferenced data demonstrating that we selected the largest values as anomalies. This is good as it is the largest values that affect the performance of a Big Data system.

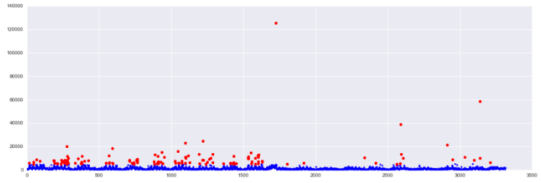


Fig. 14. DBC Workload – Average Response Time

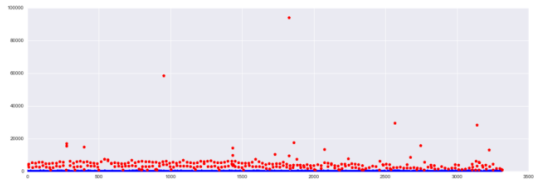


Fig. 15. DBC Workload –Total CPU Time

Confusion matrices from the both the naïve (non-feature selected) and feature selected logistic regressions run against DBC Workload data are shown in Figure 28. Generally running the feature selection and removing some of the lagged predictor variables improved the quality of the logistic models. Figure 29 shows which 10 independent lag-variables were the most frequently selected across all Workloads and KPI response variables by the recursively feature selected logistic models. It is fluctuations in those variables that are most closely tied to whether the logistic regressions predict an anomaly or not. These 10 lag-variables would be a good place to begin a root cause analysis of the source of an extreme positive anomaly.

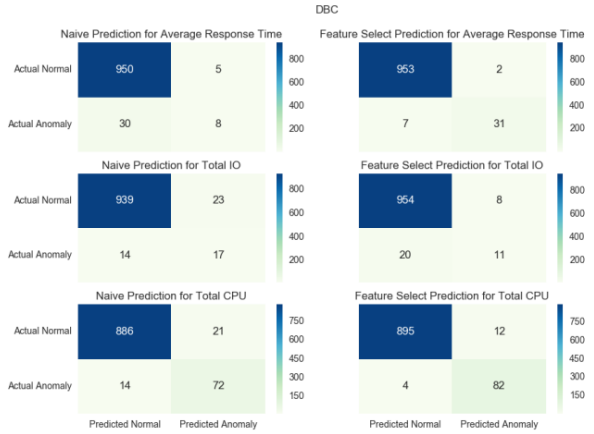


Fig. 16. Naïve and Feature Selected Logistic Regression Confusion Matrix

Top Features selected in Logistic Regression include Total Time Delay, Total Parallel Sessions, Average Response Time,

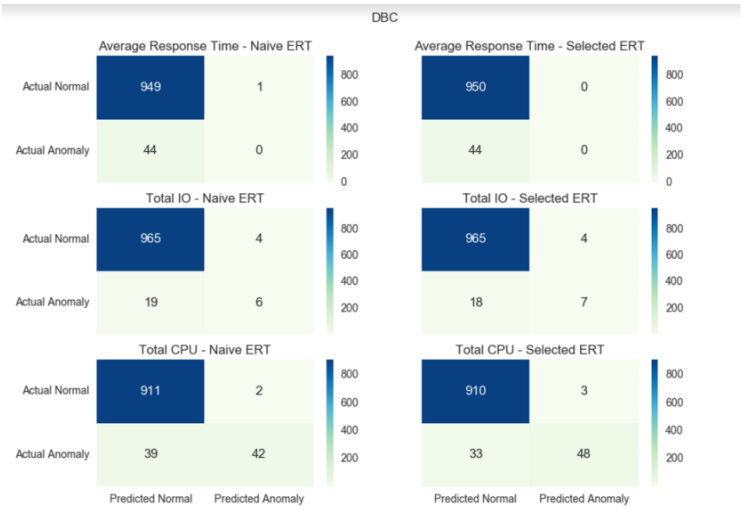


Fig. 17. Naïve and Feature Selected Extremely Randomized Trees Confusion Matrices

Top Features Selected in Extremely Randomized Trees include Total I/O, Total Priority Index, Total Intercon Messages

Forecast Accuracy:

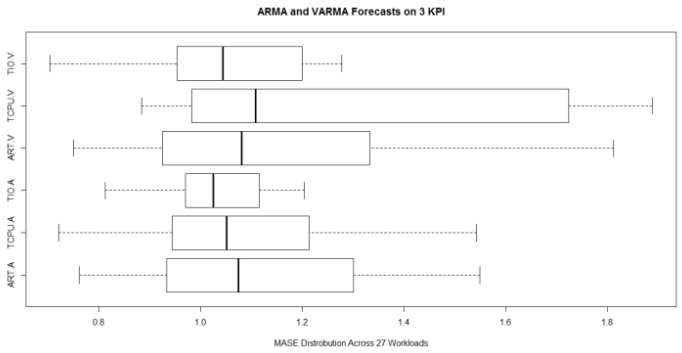


Fig. 18. Overall Forecasting MASE Values

The accuracy results were mixed with a roughly 60/40 split between models that performed worse than a naïve forecasting and models that performed better. The ARMA models tended to outperform the VARMA with 67% of the ARMA models scoring a lower MASE value than the VARMA models.

*Overall Accuracy for each Key Performance Indicator:*The Extremely Randomized Trees generally do better than the logistic regression at predicting the anomalies though it is harder to tell what the relevant variables are. The accuracy rate tables demonstrate that the Naïve (non-feature selected) ERT model generated the most accurate results at the highest rate.

Conclusion. Performance Assurance algorithms reduce uncertainty and risk of performance surprises during design, implementation and performance management of Big Data applications. Algorithm of the short term prediction based on time series is applied to identification of current anomalies and future performance problems. Diagnostic and root cause analysis reduces the scope of analysis and enables organization of the proactive performance management. Determination of the seasonality of performance characteristics improves accuracy of recommendations. Data collection, workload aggregation is performed every hour. Models are retrained every day and short term prediction algorithms are applied every hour to predict anomalies and develop proactive recommendations.

In general, the forecasting models were unable to outperform the naïve forecast. The VARMA models struggled with low forecasting accuracy and long training times. Anomalies created extreme outliers in many Workloads which were nearly impossible to forecast and caused the VARMA models to generate inaccurate outlier forecasts. Some Workloads seemed to contain additive outliers that impacted the results.

Cointegration and outliers are present in some of the Workloads which negatively impact the VARMA models. Switching to a Vector Error Correction Model might be an appropriate way to address cointegration. The anomaly detection methodology outlined in this paper or the approach described by Chen & Liu (1993) for identifying additive outliers in time series data could be used to detect significant outliers. Smoothing, or removing, the outliers may yield before training may also improve the VARMA model forecasting.

The multiple models trained for anomaly prediction provide fairly accurate results. The main issue is that there are too many false negatives, instances where an anomaly occurs and it was not predicted. The number of false negatives can be addressed by lowering the probability threshold for an anomaly but would come with a corresponding increase in false positives. The most commonly selected lagged predictor variables would be a good place to begin an analysis into what causes an extreme positive anomaly.

Two additional time series techniques may also improve the validity of the short-term predictions. First, there are methods to deal with the additive outliers that may be affecting the ARMA and VARMA fits and forecasts. Second, the cointegration results indicate that a Vector Error Correction Model may be more appropriate than a VARMA for some of the Workloads and KPIs.

Future Work. This paper focuses on implementation of the time series algorithm for short term prediction, diagnostic and root cause analysis. We are planning to apply similar approach to Performance Engineering focusing on new Big Data applications and development a Prescriptor generating recommendations related to selection of the ML algorithms and ML libraries and dynamic performance management of Big Data environment.

First, the use of a smoothing algorithm may eliminate the extreme value anomalies. The smoothing can be applied to the predictor variable time series before it is used in the VARMA model. This would prevent sudden spikes in the predictor time series from passing into the response variable prediction.

Additionally, it would be interesting to see how well a neural network could learn and predict the Workload performance. The month, day and hour previously encompassed by the time series nature of the data could be transformed into dummy variables. Pairing the dummies with the values of the Workload interaction variables would create a dataset which could be fed to a neural network. The neural network would be able to handle the high dimensionality of the data given a sufficiently large sample size.

Limitation of the common time series algorithm is a requirement not to have gaps, and Have equally spaced data. Nonlinear Time series and other algorithms providing high accurate results and do not having strict requirements to data can be evaluated. BUT from the practical point of view common time series algorithms provide acceptable accuracy and relatively fast solutions . If necessary training can be done more frequently

Finally, both the short-term forecasts and anomaly predictions can be included into a framework for monitoring and maintaining Big Data systems. Using the outputs from these models it is possible to proactively adjust resource allocation and system settings to prevent downtime or significant drops in performance.

References

[1]. Aboagye-Sarfo, P., Mai, Q., Sanflippo, Frank M., Preen, David B., Stewart, Louise M., Fatovich, Daniel M. (2015). A comparison of multivariate and univariate time series approaches to modelling and forecasting emergency department demand in Western Australia. Retrieved on October 16, 2016 from https://www.deepdyve.com/lp/elsevier/a-comparison-of-multivariate-and-univariate-time-series-approaches-to-bwna7dWZny?#

[2]. Fernández-Ros, M., Parra, J. A., Salvador, R. M., & Castellano, N. N. (2016). Optimization of the periodogram average for the estimation of the power spectral density (PSD) of weak signals in the ELF band. Measurement, 78, 207-218. doi:10.1016/j.measurement.2015.10.006

[3]. Gałecka-Burdziak, E. (2016). Aggregate matching in Spain. Time series analysis using cointegration techniques. CONTEMPORARY ECONOMICS, 10(1), 5-12. doi:10.5709/ce.1897-9254.194

[4]. Gu, X., & Wang, H. (2009). Online Anomaly Prediction for Robust Cluster Systems. IEEE International Conference on Data Engineering. Retrieved October 15, 2016, from http://dance.csc.ncsu.edu/papers/icde09-xhhx.pdf

[5]. Honaker, J., & King, G. (2010). What to Do about Missing Values in Time-Series Cross-Section Data. American Journal of Political Science, 54(2), 561-581. doi:10.1111/j.1540-5907.2010.00447.x

[6]. Lavin, A., & Ahmad, S. (2015). Evaluating Real-time Anomaly Detection Algorithms – the Numenta Anomaly Benchmark, in 14th International Conference on Machine Learning and Applications (IEEE ICMLA ’15). Retrieved October 13, 2016, from https://arxiv.org/ftp/arxiv/papers/1510/1510.03336.pdf

[7]. Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., & Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. Journal of Big Data, 2(1). doi:10.1186/s40537-014-0007-7

[8]. Villalpando, L. E., April, A., & Abran, A. (2014). Performance analysis model for big data applications in cloud computing. Journal of Cloud Computing J Cloud Comp, 3(1). doi:10.1186/s13677-014-0019-z

[9]. Zibitsker B, Lupersolsky A, CMG 2016, "Performance Engineering for New Big Data Applications

[10]. Zibitsker B, CMG 2016, "Performance Assurance for Big Data Applications"

NATURAL LANGUAGE PROCESSING FOR ECOMMERCE

A. LOPATENKO, PhD
Director of Engineering in Recruit Institute of Technology

Recruit Institute of Technology, USA
E-mail: andrei@recruit.ai

Abstract. Abstruct: Understanding users' queries and documents describing merchandise are crucial for building eCommerce applications such as shopping search engines and recommendations systems.

Entity extraction, attribute extraction, semantic understanding and tagging, query expansion improves search quality, and help to improve conversationand revenue per session.

In this presentation key NLP technologies will be presented as they used in building eCommerce applications

Author. Andrei Lopatenko is a Director of Engineering in Recruit Institute of Technology of Recruit Holdings Co. He worked for Google where he was working on algorithms for web search, Apple Inc where he was a founding engineer for Apple Maps Search and developed search algorithms for Apple AppStore and iTunes, Walmartlabs where he led a team developing search algorithms. He obtained a PhD degree in

Computer Science from the University of Manchester, UK in 2007.

His areas of expertise are organization design, building teams, building mentorship networks, running cohesive high performance teams, performance management, scaling up organizations, deep learning, search relevance, learning to rank, data science.

HOW TO SURVIVE AND THRIVE IN THE AGE OF CREATIVE DESTRUCTION

L. KATSNELSON
Director & CTO, Emerging Technologies
IBM Analytics Platform

CTO, IBM Analytics Emerging Technologies, Canada

Abstract. There is a Chinese curse which says “May he live in interesting times.” By all accounts we live in the most interesting times. Human kind has never witnessed so much danger and uncertainty. At the same time these are the most creative times in the history of mankind. This phenomenon of Creative Destruction will have its victims but it will also create many winners. The winners will be those who can gather enormous amounts of data, harness the computing power to get insights out of the data and create augmented intelligence that matches or even exceeds human intelligence.

Bio.

Leon Katsnelson, CTO, IBM Analytics Emerging Technologies

Leon has been working with data his entire career. From his office in the IBM Canada Laboratory in Toronto he leads a team of talented Developers, Data Scientists and Data Engineers. Leon and his team are always looking for ambitious projects, focusing on technologies that are just emerging on the horizon. This is a team that would rather try impossible things and risk failure than walk a well travelled path. While Leon and his team are focused on data, they are immersed in adjacent technologies such as IoT, blockchain, cloud, DevOps etc.

ТЕХНОЛОГИИ BIG DATA В СИСТЕМАХ КОНТРОЛЯ КАЧЕСТВА МЕТАЛЛУРГИЧЕСКОГО ПРОИЗВОДСТВА

Д. Н. Гайнанов
Заведующий кафедрой «Аналитика больших данных и методы видеоанализа» Уральского федерального университета им. Ельцина Б.Н.

Д. А. Беренов
Аспирант Уральского федерального университета им. Ельцина Б.Н.

Уральский федеральный университет им. Ельцина Б.Н., Россия
E-mail: damir.gainanov@gmail.com, berenov@dc.ru

Abstract. The paper considers the application of Big Data technologies in quality control systems in metallurgical production. The concept of a technological pyramid is introduced on the basis of which an approach to solving the problem of the optimal assignment of technological routes is developed and it is shown how this approach can be used to reduce the level of manufacturing defects. The proposed methods are aproved at the «Severstal» metallurgical plant within the framework of the AS SPC project (automated system for statistical production control).

Введение. В настоящее время аналитика больших данных (Big Data) находит всё большее применение в организации различных производств, в оптимизации технологических процессов. При этом важными особенностями данных, используемых для анализа, являются их объёмы (измеряемые многими терабайтами информации), потоковый характер данных (непрерывная генерация данных) и большое разнообразие используемых параметров (тысячи и десятки тысяч параметров). В приложении к металлургическому производству данный подход предполагает организацию сквозного сбора технологических параметров в процессе движения производимой продукции вдоль технологического маршрута. В результате такого процесса фиксируется история создания каждой единицы продукции с учетом прохождения её через все технологические переделы и с привязкой всех важных технологических параметров, влияющих на качество производимой продукции. В работе развивается подход к анализу данных для такой обширной базы исторических технологических данных с целью повышения эффективности производства и снижения уровня брака в производстве.

Основные определения

Рассмотрим дискретное производство, состоящее из определенного числа технологических переделов, осуществляющихся на соответствующих технологических агрегатах и линиях. Например, в качестве такого производства может служить металлургическое производство.

Пусть $A = \{A_1, \ldots, A_n\}$ – совокупность технологических агрегатов, задействованных в производстве. В данной работе важным понятием будет служить понятие единицы продукции (ЕП). Под единицей продукции понимается неделимая часть выходной или входной продукции, получаемой на агрегате или технологической линии. В качестве типичных примеров единиц продукции можно привести сталь, полученную в сталеплавильном агрегате и выпущенную в сталеразливочный ковш; слябы, получаемые после машины непрерывной разливки

стали; горячекатаные рулоны, получаемые как конечный результат работы стана горячей прокатки; холоднокатаные рулоны, получаемые как конечный результат работы стана холодной прокатки.

Определение. Ориентированный граф $G = (\mathrm{A}, E)$ с множеством вершин A и множеством дуг $E \subseteq \mathrm{A}^2$ будем называть инфраструктурным графом, если $(A_1, A_2) \in E$ тогда и только тогда, когда выходная ЕП агрегата A_1 может служить входной ЕП для агрегата A_2.

Например, для металлургического производства единица продукции сляб является выходной ЕП для машины непрерывного литья заготовок и одновременно входной ЕП для стана горячей прокатки.

Технологическим маршрутом $P = \left(A_{i_1}, A_{i_2}, \ldots, A_{i_k}\right)$ будем называть любой ориентированный путь в графе G. Множество всех технологических маршрутов $\mathrm{P} = \{P_1, \ldots, P_k\}$ будем называть технологической базой рассматриваемого производства. Обозначим через $\mathrm{EP} = \{ep_1, \ldots, ep_n\}$ множество всех возможных единиц продукции рассматриваемого производства. Пусть каждая единица продукции ep_i характеризуется набором параметров $\mathrm{P}_i = \{p_{i1}, p_{i2}, \ldots, p_{in_i}\}, i \in \overline{1, K}$. Тогда последовательность

$$AI_i = \left(A_{i1}, \mathrm{P}_{i1}(AI_i), \ldots, A_{is}, \mathrm{P}_{is}(AI_i)\right),$$

где $\mathrm{P}_{ij}(AI_i)$ – набор значений параметров для ep_{ij} в конкретной реализации технологического маршрута AI_i – будем называть исполненным технологическим маршрутом (ИТМ).

В результате производственной деятельности рассматриваемого производства будет сгенерировано множество исполненных технологических маршрутов на текущий момент времени t:

$$\mathrm{P}_{\text{итм}}(t) = \{AI_i \mid i = [1, q(t)]\}.$$

В процессе производственной деятельности происходит процесс непрерывного накопления ИТМ. Заметим, что накапливаемая информация имеет все признаки больших данных, а именно:

1 имеет значительные объемы, измеряемые многими терабайтами информации;

2 накопление информации происходит в потоковом режиме с большой скоростью;

3 накапливаемая информация характеризуется большим разнообразием и содержит значения нескольких

Определение. Технологической пирамидой $\mathrm{Pir}(G, v)$ в графе G с корневой вершиной v будем называть подграф графа G, порожденный множеством вершин $\langle v \cup G(v) \cup G^2(v) \cup \ldots \cup G^k(v) \rangle_G$ такой, что любой ориентированный путь P графа G, начинающийся в вершине v, целиком лежит в этом подграфе.

Здесь через $G^k(v)$ обозначено множество всех вершин v' графа G таких, что существует простой путь из вершины v в вершину v' длины $k-1$. Терминальной вершиной графа (подграфа) называется вершина, из которой не выходит ни одной дуги, лежащей в этом графе (подграфе). Пусть V' – множество терминальных вершин подграфа

$\langle v \cup G(v) \cup G^2(v) \cup \ldots \cup G^k(v) \rangle_G$. Для каждой терминальной вершины $v_i \in V'$ существует некоторая единица продукции ep_i, являющаяся выходной ЕП для этой вершины, причем таких ep_i может быть несколько в зависимости от вида ИТМ, в результате которого была получена данная единица продукции.

Исполненный технологический маршрут $AI_i = \left(A_{i1}, \mathrm{P}_{i1}, \ldots, A_{is}, \mathrm{P}_{is}\right)$ будем называть продуктовым, если единица продукции на выходе терминальной вершины A_{is} ИТМ AI_i – обозначим эту вершину $\mathrm{term}(AI_i)$ – является одним из видов конечного продукта, поставляемого на рынок. Любую вершину $v' \in \left(v \cup G(v) \cup G^2(v) \cup \ldots \cup G^k(v)\right)$ будем называть развилкой (рис. 1), если $\left|G\left(v'\right)\right| > 1.$

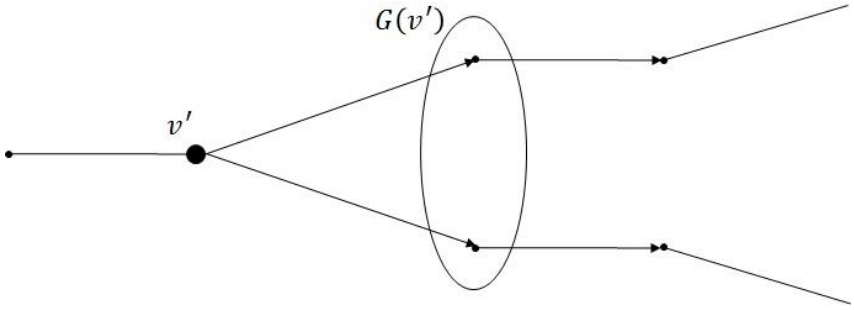


Рис. 1. Вершина–развилка в инфраструктурном графе

Контроль качества продукции на основе технологий Big Data. Предлагаемый подход может быть эффективно применен при построении системы контроля качества продукции для сложного металлургического производства. Ниже рассмотрим задачу прогнозирования брака ЕП при выполнении заданного технологического маршрута. Очевидно, что в этом случае даже такое простое решение, как прекращение дальнейшего исполнения технологического маршрута, которое может привести к производственному, браку выглядит целесообразным и, как минимум, даёт экономию на затратах на продолжении работ на последующих агрегатах, которые могут достигать существенных значений.

Для каждой ЕП, после верификации ее качества, можно определить множество наборов параметров корневой вершины, при которых получалось бы то или иное качество ep. Это будут классы в обучающей выборке. Соответственно, нам необходимо определить решающее правила в вершинах–развилках так, чтобы на обучающей выборке они правильно детерминировали вершины. В этом случае для каждого применяемого технологического маршрута $P = \left(A_{i_1}, A_{i_2}, \ldots, A_{i_k}\right)$ формируется множество ИТМ на текущий момент времени t:

$$\mathrm{P}_{\text{итм}}(t) = \{AI_i \mid i = \overline{1, q(t)}\}.$$

Для каждого ИТМ из $\mathrm{P}_{\text{итм}}(t)$ определены параметры конечной ЕП, выходящей после завершения выполнения этого ИТМ. На основе этих параметров может быть определено качество полученной единицы продукции. В простейшем случае это могут быть два класса: годные или бракованные единицы продукции.

Заметим, что для любой вершины из маршрута P, являющейся вершиной–развилкой в

технологическом графе $G=(\mathrm{A},E)$ может быть составлена обучающая выборка, состоящая из двух классов: годные и бракованные единицы продукции.

$$\begin{cases} AI_1=(\mathrm{Ind}_1, i_1, A_1, \mathrm{Pr}_{11}, A_2, \mathrm{Pr}_{12}, \ldots, A_k, \mathrm{Pr}_{1n}), \\ AI_2=(\mathrm{Ind}_2, i_2, A_1, \mathrm{Pr}_{21}, A_2, \mathrm{Pr}_{22}, \ldots, A_k, \mathrm{Pr}_{2n}), \\ \ldots\ldots\ldots \\ AI_k=(\mathrm{Ind}_k, i_k, A_1, \mathrm{Pr}_{k1}, A_2, \mathrm{Pr}_{k2}, \ldots, A_k, \mathrm{Pr}_{kn}), \end{cases}$$

где $(A_1, A_2, \ldots, A_k)$ – последовательность вершин технологического графа $G=(\mathrm{A},E)$, предшествующих вершине–развилке в маршруте P, и Ind_i – индикатор, принимающий значение 0, если единица конечной продукции при данном ИТМ получилась бракованной, и значение 1, если единица конечной продукции при данном ИТМ получилась годной. Таким образом, в этом случае получается классическая задача обучения с учителем, в которой обучающая выборка разбита на два класса.

Пусть n – совокупное множество параметров для части ИТМ маршрута P, предшествующей рассматриваемой вершине–развилке v'. Тогда обучающая выборка в данной задаче может быть записана как два множества B_0 и B_1 n-мерных векторов:

$$(\mathrm{B}_0)\begin{cases} (a_{11}, \ldots, a_{1n}), \\ \ldots \\ (a_{m1}, \ldots, a_{mn}), \end{cases}$$

$$(\mathrm{B}_1)\begin{cases} (a_{(m+1)1}, \ldots, a_{(m+1)n}), \\ \ldots \\ (a_{l1}, \ldots, a_{ln}), \end{cases}$$

где совокупность B_0 – представляет бракованные единицы продукции, и B_1 – годные единицы продукции. Тогда задача распознавания образов в геометрической постановке будет сводиться к решению следующей системы линейных неравенств:

$$\begin{cases} a_{11}\cdot x_1+\ldots+a_{1n}\cdot x_n>0, \\ \ldots \\ a_{m1}\cdot x_1+\ldots+a_{mn}\cdot x_n>0, \\ a_{(m+1)1}\cdot x_1+\ldots+a_{(m+1)n}\cdot x_n<0, \\ \ldots \\ a_{l1}\cdot x_1+\ldots+a_{ln}\cdot x_n<0. \end{cases} \qquad (1)$$

Если система совместна и $\overline{x}=(x_1, \ldots, x_n)$ – её решение, то практическое использование этого решения состоит в том, что для вновь рассматриваемого технологического маршрута при достижении вершины–развилки v' будут иметься конкретные значения параметров $(a_{j1}, \ldots, a_{jn})$, реализовавшиеся при данном маршруте. В этом случае можно прогнозировать

результат исполнения намеченного технологического маршрута P: если $a_{j1} \cdot x_1 + \ldots + a_{jn} \cdot x_n > 0,$ то прогнозируется брак конечной единицы продукции, если же $a_{j1} \cdot x_1 + \ldots + a_{jn} \cdot x_n < 0,$ то прогнозируется получение годной единицы продукции при дальнейшей реализации рассматриваемого технологического маршрута $P.$ Если же система несовместна, то для решения задачи распознавания в геометрической постановке могут быть применены методы распознавания образов, разработанные в работах [1–7].

Для заданного технологического маршрута P указанная задача может решаться для каждой вершины–развилки графа $G = (\mathrm{A}, E),$ входящей в маршрут $P.$ Таким образом, при достижении каждой вершины–развилки мы можем вычислить прогноз по тому, будет ли получено годное изделие или бракованное изделие при реализации данного технологического маршрута. Предположим, что рассмотрены все используемые технологические маршруты P и для каждой вершины–развилки технологического графа $G = (\mathrm{A}, E)$ решены задачи прогнозирования результата дальнейшего исполнения технологического маршрута. Тогда стратегия повышения эффективности производства на основе рассматриваемой в данной работе оптимизации технологических процессов на основе технологий Big Data состоит в следующем.

Пусть подлежит реализации технологический маршрут $P = (A_{i_1}, A_{i_2}, \ldots, A_{i_k}),$ в котором содержится несколько вершин–развилок $(A_{j_1}, A_{j_2}, \ldots, A_{j_m})$ технологического графа $G = (\mathrm{A}, E).$ Запускаем процесс производства до первой вершины–развилки A_{j_1} включительно:

$$(A_{i_1}, A_{i_2}, \ldots, A_{j_1}).$$

По завершении технологической операции на агрегате A_{j_1} фиксируется реализовавшийся набор параметров $(a_1, a_2, \ldots, a_{n_1}).$ Для данного маршрута P, вершины--развилки $v' = A_{j_1}$ нами ранее было вычислено решающее правило

$$\mathrm{R}(a_1, a_2, \ldots, a_{n_1}) \in \{0, 1\},$$

которое дает прогноз по конечной продукции при дальнейшем продолжении маршрута $P.$

Если $\mathrm{R}(a_1, a_2, \ldots, a_{n_1}) = 1,$ то мы продолжаем дальнейшую обработку продукции согласно технологическому маршруту до достижения следующей вершины–развилки $A_{j_2}.$ Если $\mathrm{R}(a_1, a_2, \ldots, a_{n_1}) = 0,$ то дальнейшее следование по технологическому маршруту нецелесообразно, поскольку в результате его исполнения ожидается получение брака. В этом случае необходимо проверить, существуют ли другие технологические маршруты $P_i,$ для которых уже исполненный на текущий момент времени подмаршрут маршрута $P,$ включающий текущую вершину--развилку, также является подмаршрутом маршрута $P_i.$ Если таких маршрутов нет, то дальнейшую обработку полученной единицы продукции следует прекратить и отправить данную единицу продукции на повторное использование в качестве исходного сырья. Если же

такие маршруты существуют, то необходимо для каждого такого маршрута P_i взять соответствующее правило R_{P_i} и выбрать те из них, для которых

$$\mathrm{R}_{P_i}\left(a_1, a_2, \ldots, a_{n_1}\right) = 1.$$

Далее, среди таких маршрутов следует отобрать наиболее эффективные и востребованные, возможно, с учетом текущих планов производства.

Заключение. В работе рассмотрены вопросы построения системы контроля качества в сложном многопередельном производстве, например, металлургическом с помощью методов аналитики больших данных на основе накапливаемых исторических технологических данных. Получены следующие основные результаты:

Введено понятие технологической пирамиды для инфраструктурного графа производства, играющее ключевую роль в предлагаемых методах оптимизации технологических параметров;

Приведена постановка задачи контроля качества продукции на основе технологических пирамид для системы назначения и выполнения технологических маршрутов;

Предлагается методология решения задачи назначения технологических маршрутов на основе методов распознавания образов с целью снижения брака производства;

Предлагается применение разработанной методологии оптимизации задачи технологических процессов к задаче прогнозирования и снижения брака в производстве.

Литература

[1]. Гайнанов Д. Н. Комбинаторная геометрия и графы в анализе несовместных систем и распознавании образов. М.: Наука, 2014.

[2]. Gainanov Damir. N. Graphs for Pattern Recognition. Infeasible Systems of Linear Inequalities. DeGruyter, 2016.

[3]. Гайнанов Д. Н. О комбинаторных свойствах несовместных систем линейных неравенств и выпуклых многогранников // Математические заметки. 1985. Т. 38. № 3, С. 463–474.

[4]. Мазуров Вл. Д. Метод комитетов в задачах оптимизации и классификации. М.: Наука, 1990.

[5]. Мазуров Вл. Д., Хачай М. Ю. Комитеты систем линейных неравенств // АиТ. 2004. № 2. С. 43–54.

[6]. Хачай М. Ю. Об оценке числа членов минимального комитета системы линейных неравенств // ЖВМиМФ. 1997. Т. 37. № 11. С. 1399–1404.

[7]. Gainanov D. N. Alternative Covers and Independence Systems in Pattern Recognition // Pattern Recognition and Image Analysis. 1992. V. 2. No. 2. P. 147–160.

[8]. Gainanov D. N., Matveev A. O. Lattice Diagonals and Geometric Pattern Recognition Problems // Pattern Recognition and Image Analysis. 1991. V. 1. No. 3. P. 277–282.

ВСЕЛЕННАЯ ОБЩЕСТВЕННЫХ ФИНАНСОВ. КАК АНАЛИТИКА БОЛЬШИХ ДАННЫХ МОЖЕТ ПОМОЧЬ В УПРАВЛЕНИИ ГОСУДАРСТВОМ

А. Смирнов
Hadoop-специалист Teradata

Teradata, Россия
E-mail: alexander.smirnov@teradata.com

Abstract. Современные технологии визуализации позволяют получить совершенно новые представления на основе данных, с которыми мы сталкиваемся в повседневной жизни и в работе. Они более богаты, красочны, информативны и интерактивны. Например, технологии Teradata Aster позволяют обнаружить новые закономерности, облегчают анализ и помогают в принятии решений. Вселенная открытых данных продолжает расширяться. Данные Минфина России, раскрываемые на Едином портале бюджетной системы Российской Федерации, а также данные государственных закупок и другие данные - это богатейшая почва для комплексного анализа.

ОБЛАЧНАЯ ПЛАТФОРМА SAP CLOUD PLATFORM (SCP)

А.В. Танкевич
Директор, компания JET BI
(ООО "ДЖЕТ Би Ай")

CTO, IBM Analytics Emerging Technologies, USA

Abstract. Архитектура и сервисы, предоставляемые платформой. Собственная разработка Приложений в облаке. Собственный пример разработки решения JET City Cloud.

Краткая информация о компании:

JET BI - консалтинговая компания, предоставляющая услуги в области создания и поддержки интеллектуальных решений на платформах SAP и Salesforce. Отраслевая экспертиза, выделенные команды разработчиков под каждый проект и сертифицированные специалисты гарантируют высокое качество наших услуг. Индивидуальный подход к каждому проекту - это залог эффективного решения именно для Вас.

Краткая биография. Андрей Танкевич является директором компании, со стажем работы в ИТ-сфере более 17 лет, из них 8 лет работы в качестве ИТ-директора, 5 лет – в должности директора департамента BI направления, опыт внедрения более 10 крупных проектов по разработке систем Business Intelligence.

АНАЛИЗ ЭФФЕКТИВНОСТИ ВЕДЕНИЯ БИЗНЕСА С ПОМОЩЬЮ MICROSOFT POWERBI

В. Дубовцев
Руководитель продуктивной группы Cloud+Enterprise региона стран СНГ, Республика Беларусь

ООО «Активные технологии», Республика Беларусь

СЕКЦИОННЫЕ ДОКЛАДЫ

BENCHMARKING THE EFFICIENCY OF DEEP LEARNING METHODS ON THE PROBLEM OF PREDICTING SUBJECTS' AGE BY CHEST RADIOGRAPHS

V. KOVALEV, PhD
Head of the Laboratory of Biomedical Images Analysis

V. LIAUCHUK
Research Assistant of the Laboratory of Biomedical Images Analysis

A. KALINOVSKY
Research Officer of the Laboratory of Biomedical Images Analysis

A. SHUKELOVICH
Junior Scientist of the Laboratory of Biomedical Images Analysis

Biomedical Image Analysis Department, United Institute of Informatics Problems, National Academy of Sciences of Belarus, Republic of Belarus
E-mail: vassili.kovalev@gmail.com

Abstract. This paper presents results that were obtained in comparative study of the efficiency of conventional and Deep Learning methods on the problem of predicting subjects' age by their chest radiographs. A large study group consisting of chest radiographs of 10 000 people was created by random sub-sampling of suitable subjects from the input image repository containing 1.8 million items. The age range was chosen to span from 21 to 70 years. The age prediction was performed by Convolutional Neural Networks AlexNet and GoogLeNet as well as using conventional methods based on Local Binary Patterns and extended co-occurrence matrices as image features followed by kNN, Random Forest, Linear Model, SVM, and Decision Trees classifiers. The conclusion was that the convolutional neural networks greatly outperform conventional methods. It was found that the lowest RMSE error achieved on the task of age prediction using convolutional networks is 5.77 years whereas conventional methods demonstrate on the same data much higher error value of 11.73 years.

The purpose. Recent achievements in biomedical image classification using Deep Learning methods and Convolutional Neural Networks (CNN) give well-grounded promises to become an effective tool in biomedical image analysis [1-4]. Several studies accomplished by authors on the use of CNNs for histology image classification in breast cancer diagnosis [5], lung segmentation [6] and lung lesion detection in computed tomography images of tuberculosis patients [7] confirm the applicability and power of Deep Learning methods in medical imaging domain.

In the context of a difficult choice of the most efficient machine leaning methods and software solutions for medical image analysis the primary goal of this study was to examine abilities of CNNs and to compare them to conventional methods on a large sample of chest radiographs acquired from as many as 10 000 people. The performance comparison was accomplished on the hard problem of predicting patient's age based on their chest X-ray images. Such an examination was performed using both machine learning modes including classification and regression.

Image data. A large database of natively digital chest radiographs containing about 1.8 million items resulted from pulmonary screening of population of a large city was used as input image data repository of this study. Subjects' age was measured in complete years with the precision of one year. A study group consisting of chest radiographs of 10 000 subjects was created by random sub-sampling of suitable subjects from the input image repository. The age range was chosen to span from 21 to 70 years.

In order to create a study group which is well balanced by both age and gender, for every year of life we selected 100 male and 100 female subjects what finally constituted a study group consisting of (100+100) * 50 years = 10 000 subjects. No attention has been given to the subjects' health status.

This particularly means that the created study group mostly represents healthy subjects. Nevertheless, it is still possible that a small fraction of people with certain lung abnormalities at their early stage as well as subjects with some anatomical deviations could be presented in the study group too.

Since the primary goal of this study was not the analysis of chest radiographs as such but benchmarking of Deep Learning methods, original images were preprocessed to avoid unnecessary variability of the image content and to reduce computational expenses. The preprocessing included visual quality assessment, normalization of intensity, and reformatting. The normalization of image intensity was done using commonly known technique of intensity quantiles.

Fig. 1. Examples of chest images used in this study

More specifically, a small fraction of 1% of minimal and maximal values of intensity histograms was saturated and the resultant intensity range was rescaled down to the 0-255. The image crop was performed by cutting off 25% of rows of original image size from the bottom and 5% from the other three sides. Finally, all the images were resized down to 256x256 pixels. Example images of subjects of different age and gender are presented in Fig. 1.

Experimentation outline. At the preliminary stage of preparing experimentation the input images were shuffled within every age year of each gender, i.e. within of each 100 male and 100 female subgroups of every complete year of life. Since there was sufficient amount of image data available, it was decided to subdivide the whole set of 10 000 images into the training and validation sets in the proportion of 70% to 30%. Thus, the training and the test sets consisted of 7000 and 3000 images

respectively. Once created, exactly the same training and validation image sets were used in all the experiments performed in this work. Such a technique guarantees that the results of different experiments are kept comparable in all over the study.

It should be noted that in this work we considered the subjects' age prediction results obtained on the validation image set only. This is because analysis of corresponding results achieved on the training set is typically performed for studying the issues related to the convergence characteristics, influence of certain imbalance or lack of objects of certain classes, solving the problem of overfitting, exploring the necessity of image data augmentation for a proper CNN training, etc. However, all these problems are either atypical for present work or lie outside of the scope of this paper.

Deep Learning methods. Two different approaches were used for prediction of subjects' age based on Deep Learning tools. The first approach makes use of CNNs for a direct age prediction either in regression or in classification mode. In case of classification CNN categorizes an image into one of 50 age classes each of which corresponds to 50 full age years in the range of 21-70. However, the characteristic feature of the first approach is that the final, fully-connected layer of CNN is used as a classifier.

The second approach predicts subjects' age in a similar way. The exception is only that here CNN is employed only for creation of image descriptor. The last pooling layer generated by CNN which contains 1024 elements is used as image descriptor. This layer precedes the fully-connected layer of CNN. The fully-connected layer can be viewed as an "internal" classifier of CNN and which is not used in the second approach. Instead, once created the image descriptor is extracted from CNN and supplied to an "external" classifier which could also be executed in regression and classification mode.

Combination of two training options either in regression or in classification mode and the usage of 6 different classifiers employed in this study resulted in 12 different CNN-related algorithms being examined. The list of classifiers includes internal CNN classifier (i.e., the fully-connected network) along with such external classifiers as kNN, Random Forests, Linear Model, SVM, and Binary Regression Decision Tree. These algorithms are enumerated in Tab. 1 and abbreviated for brevity. Note that the leading letter "D" stands for image descriptor created by CNN on the prediction stage.

Table 1. Twelve age prediction algorithms utilizing CNNs and their abbreviations.

No	Internal/External classifier of CNN	Algorithm abbreviation (executed in 2 modes)
1	Fully Connected Layer of CNN (internal)	CNN
2	kNN (descriptor-based, external)	D-kNN
3	Random Forests (descriptor-based, external)	D-RF
4	Linear Model (descriptor-based, external)	D-LM
5	Support Vector Machines (descriptor-based, external)	D-SVM
6	Binary Regression Decision Tree (descriptor-based, external)	D-DT

Training convolutional networks. Two convolutional networks AlexNet [8] and GoogLeNet [9] were trained under Linux operating system using the Caffe framework from Berkeley Learning and Vision Center which supports GPU acceleration via cuDNN to massively reduce training time [10]. The Caffe framework was chosen from the list of freely available Deep Learning frameworks [11] because of the following reasons:

– Recently, the Caffe framework is one of the best frameworks optimized for GPU-based computing using convolutional networks for image classification.

– The Caffe framework is supported by Nvidia company and it is integrated into the Deep

Learning GPU Training System (DIGITS) interface [12] which provides high level user interface.

– The Caffe framework is well supported by a large academic community which provides voluntary consulting, share pre-trained and trained CNNs for free, etc.

The training was performed on a personal computer equipped with recent Intel® Core™ i7 central processor and two GPU of Nvidia TITAN X type with 3072 CUDA Cores and 12 GB of GDDR5 onboard memory each. The network training parameters were set to the following values:

– Network architectures: AlexNet, GoogLeNet (the first version of Inception architecture from Google).
– Batch size: 32 (the minimum batch size to place network in GPU memory).
– Solver: SGD Caffe solver.
– Number of iterations: 13 140.
– Number of epochs: 60.
– Training set size: 7000 images, 256x256 pixels each.

Age prediction with the help of internal CNN classifier was performed using DIGITS interface. Several Python scripts using PyCaffe interface were written for extracting image descriptors produced by convolutional layer of CNN and inputting them into external classifiers.

The network training time varied from 30 to 60 minutes depending on such parameters as batch size, number of iterations and some other.

Results achieved with Deep Learning methods. The first series of experiments with AlexNet and GoogLeNet convolutional networks have revealed that GoogLeNet slightly but systematically outperforms AlexNet in the quality of subjects' age prediction by chest radiographs. In terms of the Root Mean Square Error (RMSE) of the deviation of predicted age from real one the prediction quality achieved by GoogLeNet was approximately for 0.3-0.8 years better comparing to the one provided by AlexNet. Thus, all the prediction results reported below were obtained with the help of GoogLeNet.

Results of all twelve experiments measured in RMSE error presented in a condensed form on the left panel of Fig. 2.

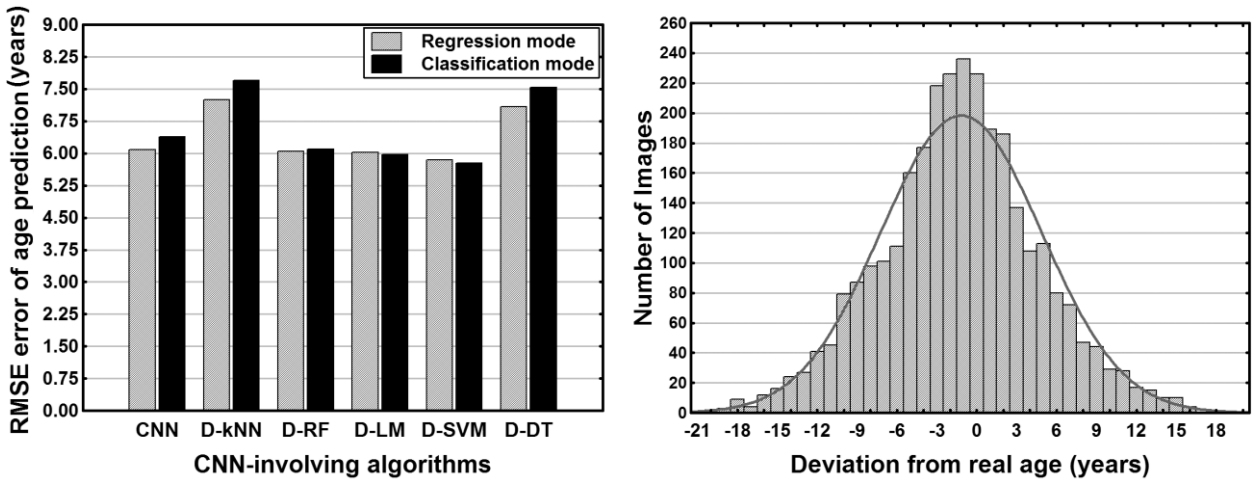


Fig. 2. Results of predicting age of 3000 subjects using convolutional network. Left panel: RMSE error for 12 different algorithms (the lower the better). Right plot: example histogram of residuals.

The right panel provides rather typical example of histogram of residuals in case of running GoogLeNet in regression mode with the native fully-connected layer as a classifier (see the first bar of the left plot).

Local Conclusions. Results of age prediction presented in Fig. 2 allows drawing the following local conclusions.

– Depending on the specific algorithm employing CNN the mean error of age prediction varies in the range from 5.77 years in the best case using of image descriptors created by GoogLeNet in classification mode which were inputted to the external SVM classifier up to 7.25 years of mean error for the same descriptors supplied to kNN classifier.

– The "direct" age prediction by CNN (i.e., without additional manipulation with extracting image descriptors and employing external classifiers) do not provide the best results. However, it is reasonably good with its RMSE value of 6.08 (see the first column of the left plot of Fig 2) compared to the best one of 5.77 achieved by CNN followed by SVM.

– Despite the best value obtained in case of running CNN in classification mode, there can be some tendency observed for better results being achieved when CNN is used in regression mode (see gray bars in Fig. 2). The reason behind could be purely technical such as classification for 50 age classes provide integer age output whereas regression predicts age with the precision of a fraction of year.

The histogram of residuals depicted on the right panel of Fig. 2 demonstrates relatively good fit to the Gaussian distribution what is suggestive for bias-free prediction model.

Conventional methods. Prediction of subjects' age based on chest X-ray images was done by implementation of a three-step procedure comprising of calculation of image descriptors, performing the Principal Component Analysis (PCA) and inputting resultant features into classifiers for age prediction. Below these steps are described in more details.

Step-1: Calculating image descriptors. Since chest images used in this study exhibit typical textural appearance, texture features were employed as image descriptors. Two kinds of texture features were used in order to obtain more extensive, reliable and trustful results. They include commonly known Local Binary Patterns (LBP, [13]) and extended multi-sort and multi-dimensional co-occurrence matrices introduced in [14].

In case of LBP rotation-invariant versions of both uniform and non-uniform types of binary patterns were examined. In case of co-occurrence image descriptors we used 2D version of six-dimensional co-occurrence matrices [15] abbreviated as IIGGAD which fuse intensity (denoted by I) gradient magnitude (G) and anisotropy (angle between gradient vectors A) image features for pixel pairs with inter-pixels distances ranged from 1 to D. It is easy to see that the classical intensity co-occurrence matrices IID with varying inter-pixel spacing can be viewed as a reduced version of the above general case. Next, the particular case of AD type gives us some rotation-invariant version of widely used Histogram of Oriented Gradients (HOG), etc. Technically, all reduced versions can be obtained from IIGGAD by summing up (collapsing) the unnecessary dimensions. It should be remembered also that dimensionality of extended co-occurrence matrices depends on the number of selected features characterizing the pixel pair and not related to image dimensionality (see [15] for more details).

Step-2: Principal Component Analysis. The above "raw" texture features (e.g., elements of co-occurrence matrices) can be very large and contain mutually-correlated elements. Performing PCA dramatically reduce feature space and resulted in uncorrelated principal components which contain essentially the same information because any original variable can be presented as a linear combination of principal components.

Step-3: Age prediction. The principal components obtained on the previous step were considered as image features. Similar to the neural networks case considered in previous sections they were inputted into the same kNN, Random Forests, Linear Model, SVM, Binary Regression Decision Tree classifiers and executed in regression mode to predict subjects' age. In all the occasions the control parameters were kept same to make comparison of results obtained by conventional methods and CNNs straightforward.

Results achieved with conventional methods. Preliminary experiments. In context of this particular study it should be emphasized that in the case of using convolutional network there were al-

most no control parameters notably influencing the quality of image descriptors produced by convolutional layers. However, this is not the case with LBP and extended co-occurrence features. Thus, in order to avoid unnecessary favoritism towards newly immerged Deep Learning tools there were a number of experiments performed for tuning control parameters of LBP and extended co-occurrence image descriptors.

A total of 18 variants of rotation-invariant LBP descriptors were examined including 9 uniform and 9 non-uniform versions with the radius of local circular neighborhood of 1, 3 and 5 pixels and number of pixels compared to the central one equal to 8, 12 and 16. As a result it was found that depending on combination of these parameters the RMSE error varied in the range from 12.93 to 17.30 years.

Similar investigation was performed for extended co-occurrence matrices. A total of 32 variants of IIGGAD, AD, and GGD matrices were evaluated with the number of intensity bins equal to 8, 16, 24 and 32, gradient magnitude bins equal to 8 and 16, number of angle bins 12 and 16 as well as inter-pixel distances of 1, 3 and 5 pixels. Note that not all possible combinations of control parameters were tested. Also, it was found that contrary to a wide spread believe an increase of intensity resolution (number of intensity bins) towards 256 not necessarily increases quality of final results. Finally, the exploratory experiments with extended co-occurrence descriptors have revealed that RMSE error of age prediction varied between 13.12 and 16.84 years what is similar to the range obtained when using LBP.

Principal Component Analysis. Instead of selecting the most prominent variants from 18 particular LBP and 32 co-occurrences image descriptors described above they were merged into two corresponding data tables as subsets of variables and supplied to PCA. As a result we got two sets of image features obtained with the help of LBP and extended co-occurrence image descriptors. In all the occasions the output principal components were selected so that they explain 99% of variation of raw image descriptors of training image dataset. As a result, the number of selected principal components varied from several dozen up to one hundred.

It is important to note that the output principal components derived from raw LBP and extended co-occurrence image descriptors were also mutually correlated. This is not surprising because these two kinds of features describe quantitatively the content of the same image set. Finally, they were inputted together into the "second" PCA for obtaining a set of joint image features combining advantages of both LBP and extended co-occurrence image descriptors.

Final results. The results obtained based on LBP, extended co-occurrence and joint image features using 5 different classifiers are presented in Fig. 3. Similar to age prediction results obtained with convolutional networks, they measured in RMSE error. The right plot of Fig. 3 depicts histogram of residuals obtained based on the joint image descriptor using SVM classifier and illustrates proportions of different prediction errors.

Local Conclusions. Results of age prediction by the subjects' chest X-ray images using conventional methods which are presented in Fig. 3 allow to draw the following local conclusions.

– The use of joint image descriptors always provide better result comparing to LBP and extended co-occurrence alone (black bars in Fig. 3 are lower for all 5 classifiers).

– Formally, the best result with minimal error value of 11.73 years was achieved based on joint image descriptors using SVM classifier. However, this is only for 0.04 and 0.15 years better than prediction provided with the help of Linear Model and Random Forest respectively.

– The LBP descriptors alone (see gray bars) slightly outperform extended co-occurrence matrices. For instance, in case of Random Forest classifier the error value of LBP is 12.39 against 12.61 years achieved by co-occurrence what corresponds to subtle difference of 0.22 years. The highest difference of errors in age prediction of 0.60 years between these two descriptors is observed in case of kNN (16.24 vs. 16.84) whereas with Decision Tree classifier LBP and co-occurrence descriptors demonstrate same error of 15.50 years.

– Random Forest, Linear Model and SVM classifiers doing always better than kNN and

Decision Trees.

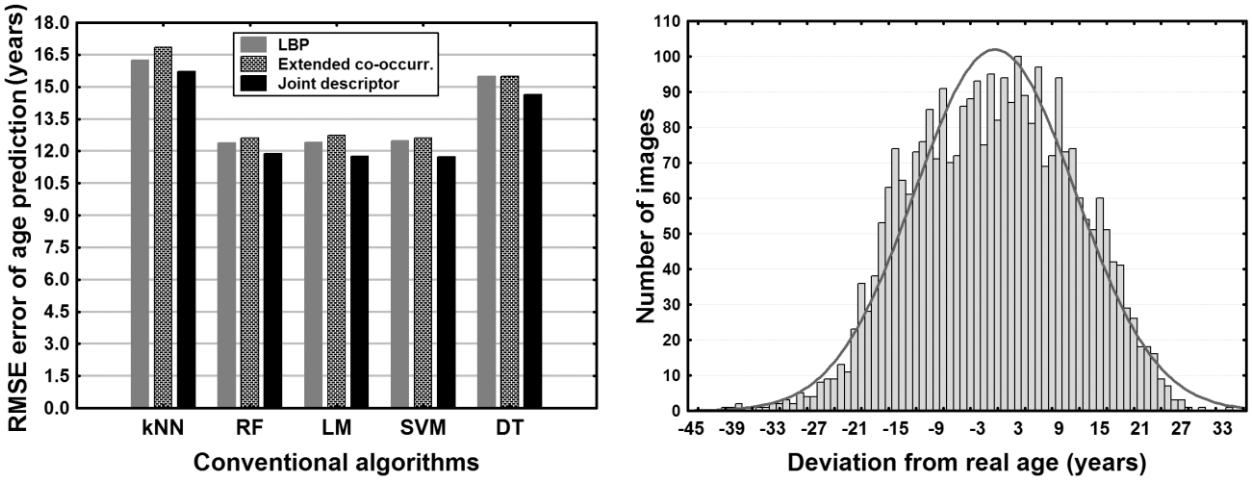


Fig. 3. Results of predicting age of 3000 subjects using conventional algorithms. Left panel: RMSE error obtained using 3 types of image descriptors and 5 classifiers (the lower the better). Right plot: example histogram of residuals.

As can be seen from the histogram of residuals which was calculated as the difference between real and predicted age, there is a tendency for overestimating subjects' age when using conventional approaches (see the shift to negative values of histogram of Fig. 3).

Conclusion. Results obtained with this comparative study of the efficiency of conventional and Deep Learning methods on the problem of predicting subjects' age by their chest radiographs allow drawing the following conclusions.

(1) The convolutional neural networks greatly outperform conventional methods. The lowest RMSE error achieved on the task of age prediction using convolutional networks is 5.77 years whereas conventional methods demonstrate on the same data much higher error value of 11.73 years.

(2) The worst error value of 7.25 years obtained in 12 experiments with neural networks is still far better than the best result of 11.73 year error obtained in 15 experiments following conventional approach. In general, results obtained with convolutional network approximately twice as good comparing to the conventional methods examined in this study.

(3) Results produced by convolutional layers during the network training can be used as compact image features describing the image content.

Acknowledgements. This study was partly supported by the National Institute of Allergy and Infectious Diseases, National Institutes of Health, U.S. Department of Health and Human Services, USA through the CRDF project OISE-16-62631-1.

References

[1]. Ravi D., Wong C., Deligianni F., Berthelot M., Andreu-Perez J., Lo B., Yang G.Z. Deep Learning for Health Informatics, IEEE Journal on Health and Biomedical Informatics, vol. 21, No 1, 2017, pp. 4-21.

[2]. Deep Learning for Medical Image Analysis, Zhou S. K, Greenspan H., Shen D. (Eds), Academic Press, ISBN: 9780128104088, 2017, 458 p.

[3]. Litjens G., Kooi T., Bejnordi B.E., Setio A.A.A., Ciompi F., Ghafoorian M., van der Laak J.A.W.M. , van Ginneken B., Sánchez C.I. A Survey on Deep Learning in Medical Image Analysis, arXiv:1702.05747, 2017, 34 p.

[4]. Kovalev V., Kalinovsky A., and Kovalev S. Deep Learning with Theano, Torch, Caffe, TensorFlow, and deeplearning4j: which one is the best in speed and accuracy? In: XIII Int. Conf. on Pattern Recognition and Information Processing, 3-5 October, Minsk, Belarus State University, 2016, pp. 99-103.

[5]. Kovalev V., Kalinovsky A., Liauchuk V. Deep Learning in Big Image Data: Histology image classification for breast cancer diagnosis, In: Big Data and Advanced Analytics, Proc. 2nd International Conference, BSUIR, Minsk, June 2016, pp. 44-53.

[6]. Kalinovsky A. and Kovalev V. Lung image segmentation using Deep Learning methods and convolutional neural networks . In: XIII Int. Conf. on Pattern Recognition and Information Processing, 3-5 October, Minsk, Belarus State University, 2016, pp. 21-24.

[7]. Liauchuk V., Kovalev V., Kalinovsky A., Tarasau A., Gabrielian A., Rosenthal A. Examining the ability of convolutional neural networks to detect lesions in lung CT images. Journal of Computer Assisted Radiology and Surgery, CARS-2017 International Congress, Barcelona, 20-25 June 2016 (in press).

[8]. Krizhevsky A., Sutskever I., Hinton G. Imagenet classification with deep convolutional neural networks, In: Advances in neural information processing systems, 3-8 December, USA, 2012, pp. 1097-1105.

[9]. Szegedy, Christian, et al. Going deeper with convolutions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1-9.

[10].Jia Y., Shelhamer E., Donahue J., Karayev S., Long J., Girshick R., Darrell T. Caffe: Convolutional architecture for fast feature embedding. In: Proceedings of the ACM International Conference on Multimedia, ACM, 2014, pp. 675-678.

[11].Kovalev V., Kalinovsky A., and Kovalev S. Deep Learning with Theano, Torch, Caffe, TensorFlow, and deeplearning4j: which one is the best in speed and accuracy? In: XIII Int. Conf. on Pattern Recognition and Information Processing, 3-5 October, Minsk, Belarus State University, 2016, pp. 99-103.

[12].https://devblogs.nvidia.com/parallelforall/deep-learning-computer-vision-caffe-cudnn/ Last visited March 2017.

[13].Pietikäinen M., Hadid A., Zhao G., Ahonen T. Computer Vision Using Local Binary Patterns. Volume 40, Springer-Verlag, London, 2011, ISBN 978-0-85729-747-1, DOI 10.1007/978-0-85729-748-8.

[14].Kovalev V. and Petrou M. Multidimensional co-occurrence matrices for object recognition and matching, Graphical Models and Image Processing, vol. 58, No. 3, pp. 187-197, 1996.

[15].Kovalev V.A., Kruggel F., Gertz H.-J., and von Cramon D.Y. Three-dimensional texture analysis of MRI brain datasets, IEEE Transactions on Medical Imaging, vol. 20, No. 5, pp. 424-433, 2001.

BUSINESS CONTEXT OF BUSINESS INTELLIGENCE (WORKSHOP LESSEN 1)

M. G. STROO, PhD
Owner of Invisi, Netherlands, Owner of Act On Insight, Belarus, Information Innovation Leader, Business Intelligence Consultant: Royal Agio Cigars, City of Rotterdam, Nuon

Invisi BV, Netherlands

Business Context Of Business Intelligence

Opening Exercise

- You will work for one of these companies during the course:
  - Belarus Tractor (production)
  - Act On Insight (IT services)
  - MTS (telecom)
- Choose your company from a raffle

Business Intelligence Definition

- Definition of BI as a process:
  Business Intelligence is the *continuous* process with which organisations can gather and register, analyse data in a structured manner and *use* the resulting information and knowledge in decision making processes to *improve* the *performance* of the organisation.

Business Intelligence Definition

- Definition of Business Intelligence as technology:
  Business Intelligence is the collection of IT resources that *supports* Business Intelligence as a process, makes it efficient and gives it a face.

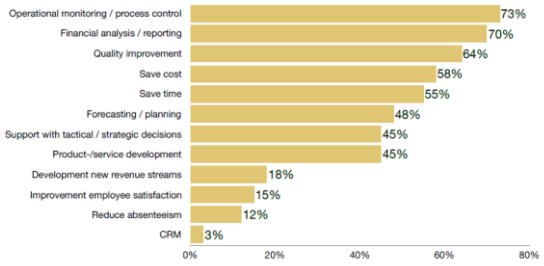
Business Intelligence Definition

- Definition of Business Intelligence as phenomenon or discipline:
  Business Intelligence is the whole of concepts, processes, strategies, culture, structure, methods, standards and IT resources that ensure that organisations *can* behave and *develop* themselves more intelligent.

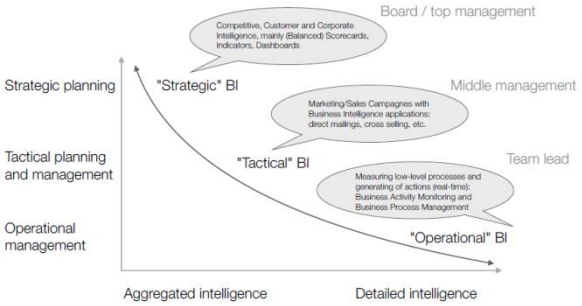
Why Business Intelligence

- "Measuring is knowing, guessing is missing"
- Make informed decisions
- Fact-based learning
- Lower cost, increase revenue
- Stabilise or improve competitiveness
- Increase customer knowledge
- Anticipate to changes in the market
- Mandatory regulations, such as:
  - Sarbanes-Oxley Act
  - Basel I-II-III
  - IFRS

Which goals does your organisation want to reach or support with BI?



Three Levels of Business Intelligence



Business Intelligence is "more than reporting"

Disciplines of Business Intelligence

- Advanced Analytics
- Management Information
- Performance Management
- Data Discovery
- Reporting and Dashboarding
- Balanced Scorecarding
- Marketing Intelligence
- Predictive Analysis
- Data Mining
- Analytical CRM
- Business Activity Monitoring

Business Intelligence cycle variations

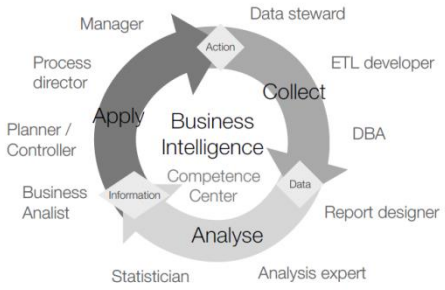
- Strategical / tactical / operational level
- Frequency of use: from yearly and ad hoc to daily and real-time
- Planning: e.g. collect daily, analyse weekly and apply monthly
- Scope: from individual and department to company and chain
- Explicitness
- Area of use

Data Warehousing

Business Intelligence

Data Warehousing

Business Intelligence is (mostly) human work



Types Of Business Intelligence Users

User Type: Farmers - Clear Sighted

- Monitor the effect of decisions on the business by tracking key performance metrics and analysing business reports
- Provide Explorers/Miners with feedback on the effectiveness of their predictions

User Type: Farmers - Clear Sighted

- Have a fairly predictable pattern of usage
  - They know what data they want, how they want it displayed, when they want it and in what media
- See the world in terms of dimensions (time, product, geography) and metrics (usage, counts, revenue, costs)

User Type: Farmers - Clear Sighted

- Farmers mainly use multidimensional data marts
- Examples of farmers:
  - Sales Analysts
  - Financial Analysts
  - Market Campaign Managers
  - Accounting Analysts

User Type: Explorers - Innovative

- Work to understand what makes the business work by looking for hidden meanings in corporate data
- Have little or no idea what to expect from query execution
  - An out of the box thinker
  - Launches large and often unpredictable queries
  - Often receives no results back
  - Occasionally receives incredible insight

User Type: Explorers - Innovative

- Strive to predict the future based on past results
- Very knowledgable about data content within and outside of the business
- Demonstrate an unpredictable pattern of usage
- Sees the world in terms of data and data relationships

User Type: Explorers - Innovative

- Explorers may start with multidimensional data marts but often require their own environment
- Examples of explorers:
  - Insurance Actuaries
  - Process Control Engineers
  - Market Research Analysts

User Type: Miners - Thorough

- Scan large amounts of detailed data looking for confirmation of a hypothesis or for suspected patterns
- Have a pretty good idea what to expect prior to query execution
- Operate on a base of data that is preconditioned for analysis

User Type: Miners - Thorough

- Demonstrate a reasonably predictable pattern of usage
- Interested in finding meaningful relationships in transactions

User Type: Miners - Thorough

- Miners may start with multidimensional data marts but often require their own environment
- Examples of Miners:
  - Expert Marketers
  - Risk Controllers
  - Logistic Specialists
  - Statisticians

User Type: Tourists - Generalists

- Have a broad business perspective and are aware of the data produced by the business
- Use the data warehouse frequently
- Cover a breadth of material quickly but in little depth
  - Are accustomed to a consistent graphical user interface
  - Need ability to search large banks of data without a lot of typing
- Demonstrate unpredictable patterns of usage

User Type: Tourists - Generalists

- Tourists mainly use multidimensional data marts and/or informal warehouses
- Examples of Tourists:
  - Executives
  - Managers
  - Casual users

User Type: Operators - Focused

- Use the intelligence derived by Explorers and Farmers to improve business conditions
- Provide increasing pressure on the Corporate Information Factory in terms of availability, data freshness and query performance
  - Need fresh, detailed, day-to-day information
  - Expect transactional performance and response times

User Type: Operators - Focused

- Have a fairly predictable pattern of usage
- See the world in terms of process

User Type: Operators - Focused

- Operators mainly use the operational data store and sometimes multidimensional data marts
- Examples of Operators:
  - Customer Support Representatives
  - Manufacturing Line Managers
  - Inventory Control Managers

User Types Exercise

- Identify two departments in your company
- Name one role in each of the departments
- Say for each role what user type it is
- Name a typical type of BI application for each role

Business Intelligence Maturity Scan

Business Intelligence Maturity Matrix

BIMM	Local	Coordinated	Integral	Intelligent
BI Architecture	- Independent Data Marts - Large tool variation - Missing standards - Limited attention to Data Quality - Reporting / limited OLAP	- Convergence of Data Warehouse and Data Marts - Standardisation of tools - BI portals - Exchange technical metadata - OLAP servers	- Hub and spoke / federated - Standardisation of methods - Data cleaning - Common metadata - Visualisation, notification, collaboration - Common BI framework	- Enterprise Information Integration / Data Bus - Continuous improvement of work processes - BI (web) services - Collaboration - Real-time closed loop - Total Data Quality Management
BI Organisation	- Various project teams mainly IT staffed - Local customers - Ad hoc development and management	- Shared project office - Program management - Management by IT department - Professionalisation	- BI Competence Center - Reuse and repeatable quality - BI Governance - Management by CIO	- Shared Service Center - BI fully part of business management and operations - BI is a board issue
BI Ambition	- Understand - Department level - Accountability afterwards	- Improve - Limited consolidation - Adjust timely	- Optimise - Integral company policy - Proactively application of information	- Innovate - BI also for external partners and customers - Intelligent organisation

BI Maturity Scan Business Questions

- How do you see the organisation?
- What are your priorities at the moment?
- How do you know it is going well?
- What are your business worries?
- How do you measure success?
- How often do you check your most important measures?
- Which applications does your department use and what do you think about them?

BI Maturity Scan Technical Questions

- Which Business Intelligence and Data Warehouse related hardware and software is in use?
- How, when and why is chosen for this setup?
- How are changes to the system chosen and applied?
- How do you receive requests to supply information or make it available for consumption?
- What is available in standards and documentation?
- Which issues are there in relation to the system?

Business Intelligence Maturity Scan Excercise

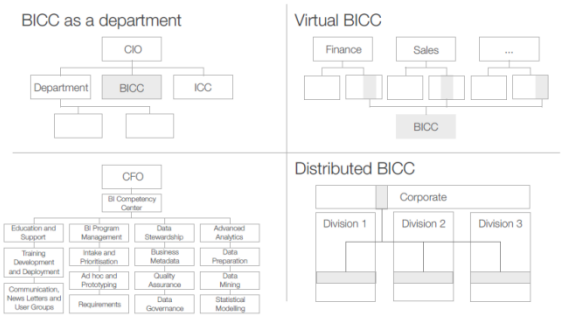
- Choose 4-7 key people from your company to interview
  - State their department and role
- Answer the top five questions of each subject area briefly from your company key people
- Draw your basic maturity matrix by writing the main factor in the appropriate block for each subject area

Business Intelligence Competency Centre

Business Intelligence Competency Centre

- A business Intelligence Competency Centre (BICC) is a permanent formal organisational structure. It includes representatives from the business and IT. Its aim is to advance and promote the effective use of Business Intelligence to support the organisation's business strategy.
- BI CC targets:
  - Ensure and share business, IT and analytical knowledge
  - Actively support BI projects
  - 'Educate' and advise management
  - Broad communication about use and necessity of BI
  - Actively develop BI knowledge
  - Recognise the importance of BI architecture
  - Have the IT department deal with the diversity of BI technologies

BICC Positioning

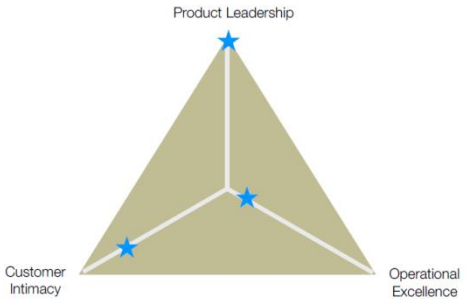


Which BICC structure to choose?

- Depends on:
  - Type of decisions
  - Available resources
  - Organisational structure
  - Organisation culture
- The following questions can help:
  - Which type plans and decisions does the intelligence have to found (marketing, financial, production, strategic plans)?
  - Which relevant resources are available and where (sources, experts, analysts)?
  - What is the relationship between the business units (same markets, departments, competitors, clients, information needs, means)?
  - Is the organisation centralised or decentralised?

Key Performance Indicators

What is the Strategy?



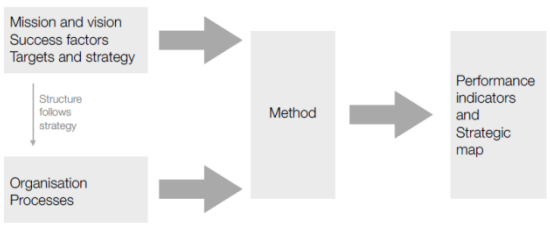
Strategy and Indicators

- Organisations use scorecards to visualise results
- A scorecard is a set of indicators grouped in perspectives
- An indicator is a number or a graph that shows the results in a certain area

Advantages of Managing With Indicators

- Indicators...
  - offer a compact representation of essential information
  - are easily interpreted through visual representation
  - give focus: aim attention to essentials
  - show performance at a glance
  - can be delivered fast
  - are flexible in design
  - give insight in trend and expected developments
  - together are more than the sum of their parts

From Strategy to Indicators



Four Different Approaches to Formulate Indicators

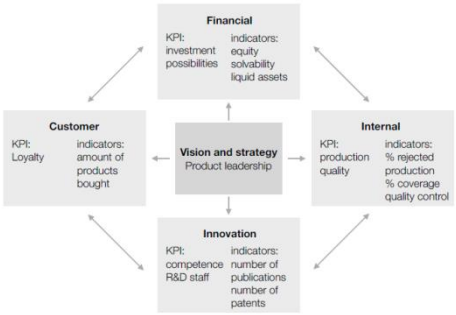
- Balanced Scorecard
- Process and Integral Chain Approach
- Horizontal Approach
- Vertical Approach

Balanced Scorecard Overview

Balanced Scorecard

- Four perspectives:
  - Financial: how do shareholders and other investors see us?
  - Customers: how do (potential) customers see us?
  - Internal: What must we excel at? How do our processes perform?
  - Innovation: How can we continue to improve, create value and innovate?

Balanced Scorecard



Balanced Scorecard

- (Strategic) objective
  - Key Performance Indicators
    - Indicators
- Limit your KPI's!
- Typical is one or two KPI's per subject area

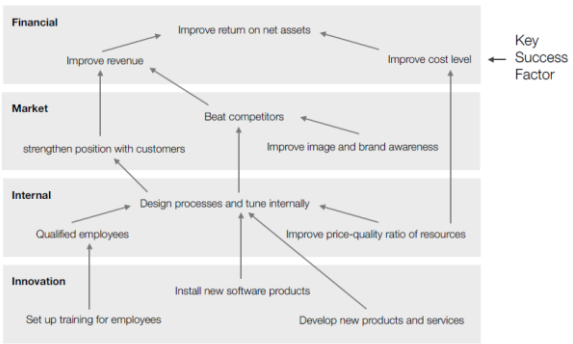
Value of the Balanced Scorecard

- Vision of Top management
- Key performance indicators
- Gives a balanced set of indicators
- Focus on continuity and controlled change
- Stimulates full control
- Suitable for professionalising of management
- Makes strategy explicit and communicable

Success Factors and Causalities

- Key Success Factors are...
  - those factors that determine the success of the unit, ultimately leading to the economical performance of that unit
  - the aspects that are crucial to realising the strategy
  - the core priorities for managing quality and productivity
- Causalities show the dependencies between Key Success Factors

Balanced Scorecard - Strategic Map Example



Balanced Scorecard - Indicators Example

- Strengthen position with customers
  - % new revenue with existing customers
  - Profit margin per customer
  - Revenue with new customers
  - Average lead time
  - Effort ratio to hours
- Improve image and brand awareness
  - Spontaneous brand recognition
  - Helped brand recognition

Master Card for an Indicator

Master Card		Activity Ratio to Hours
General	Nr	823
	Definition	Size new revenue / number of hours sales activity
	Type	€ 0
	Owner	Director Marketing and Sales
	Perspective	Internal
	Direction	Increasing
	Norm	t.b.d.
	Bandwidth	t.b.d.
	Norm setter	Director Marketing and Sales
	Report frequency	Monthly
Measurement	Supplier	Department MI administration
	Source systems	Sales system and time sheets
Known issues		Time sheets of sales representatives lags 2 weeks

Characteristics of a Good Indicator

- The indicator shows at a glance the situation of a certain aspect of the organisation or other relevant matters
- Developments in time are visible to make clear if there is a positive, negative or flat trend
- An indicator is aimed on priorities, to ensure as little attention as possible is spent on side issues
- The indicator has a norm to make clear which conditions call for unchanged ('green'), preventive ('orange') or corrective ('red') measures. The norms are ideally derived from the strategic organisation objectives
- The meaning of an indicator is clear, with only one interpretation and version of the truth. When for instance a manager and internal accountant disagree about the meaning of an indicator, this dispute has to be solved. This may lead to two indicators with a clearly distinguishable name and definition

Variation of and Alternative for Balanced Scorecard

- Internet Scorecard
  - For organisations working primarily online, or for monitoring the online channel
  - Dimensions:
    - Financial
    - Visitor
    - Website (including infrastructure)
    - Organisation
- Triple Bottom Line
  - https://en.wikipedia.org/wiki/Triple_bottom_line
  - People
  - Planet
  - Profit
- Circles of Sustainability
  - https://en.wikipedia.org/wiki/Circles_of_Sustainability
  - Economics
  - Ecology
  - Politics
  - Culture

Key Performance Indicators Exercise

- Describe one KPI in one subject area for your company
- Describe three to five indicators related to this KPI

DATA WAREHOUSE ARCHITECTURE AND DESIGN (WORKSHOP LESSEN 2)

M. G. STROO, PhD
Owner of Invisi, Netherlands, Owner of Act On Insight, Belarus, Information Innovation Leader, Business Intelligence Consultant: Royal Agio Cigars, City of Rotterdam, Nuon

Invisi BV, Netherlands

Data Warehouse Architecture And Design

Data Warehouse Architectures

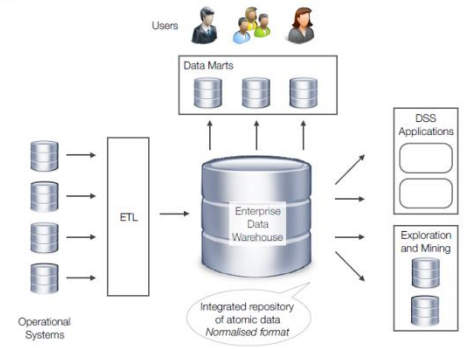
Inmon's Corporate Information Factory

- This is a hub-and-spoke architecture
- The core is a single repository called the 'Enterprise Data Warehouse'
- It is an integrated repository of atomic data:
  - Integrated from the various operational systems
  - Atomic as the data is captured at the lowest level of detail possible

Inmon's Corporate Information Factory

- The enterprise data warehouse is not intended to be queried directly by analytic applications, business intelligence tools, or the like
- Its purpose is to feed additional data stores dedicated to a variety of analytic systems
- Inmon advocates the use of third normal form database design for the enterprise data warehouse
- Inmon uses the term ETL only for the movement of data from the operational systems into the enterprise data warehouse
- He describes the movement of information from the enterprise data warehouse into data marts as "data delivery"

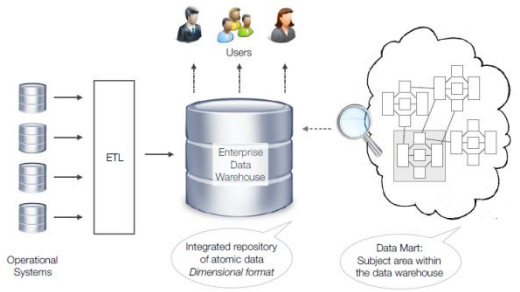
Inmon's Corporate Information Factory



Kimball's Dimensional Data Warehouse

- Kimball is largely responsible for popularising star schema design in the 1990's
- Kimball developed an enterprise architecture for the data warehouse, built on the concept of dimensional design
- Sometimes referred to as the "bus architecture"
- Shares many characteristics of Inmon's Corporate Information Factory

Kimball's Dimensional Data Warehouse



Kimball and Inmon Similarities

- Separation of the operational and analytic systems
- ETL process to consolidate, integrate and load the data into a single repository
- Data goes into an integrated repository of atomic data

Kimball and Inmon differences

- The dimensional data warehouse is designed according to the principles of dimensional modelling. It consists of a series of star schemas or cubes, which capture information at the lowest level of information possible
- The enterprise data warehouse is designed using the principles of ER (entity-relationship) modelling.
- The dimensional data warehouse may be accessed directly by analytic systems. The data mart becomes a logical distinction; it is a subject area within the data warehouse

Kimball and Inmon Variations

- An intermediate step with a set of tables in third normal form to make the ETL easier is acceptable for Kimball
- These are usually staging tables and should be accessed directly only by the ETL process
- This makes the Kimball solution more like the Inmon solution, with a normalised repository of data not accessed by applications
- Another variant is where the dimensional data warehouse is not accessed directly by analytic applications
- New data marts are constructed by extracting data from the dimensional data warehouse
- This increases the resemblance to the Corporate Information Factory, where data marts are separate entities from the integrated repository of atomic data

Stand-Alone Data Marts

- Can achieve rapid and inexpensive results in the short term
- The stand-alone data mart is an analytic data store that has not been designed in an enterprise context
- It is focused exclusively on a subject area
- One or more operational systems feed a database called a data mart
- Analytical tools or applications query it directly, bringing information to end users
- Data marts may be offered as part of packaged (operational) applications
- Sometimes they are built within user organisations, outside of the IT department

Architecture and Dimensional Design

Architecture	Advocate	Also known as	Description	Role of Dimensional Design
Corporate Information Factory	Bill Inmon	• Atomic data warehouse • Enterprise data warehouse	• *Enterprise data warehouse* component is an integrated repository of atomic data • It is *not* accessed directly • Data marts reorganise data for departmental use / analysis	Dimensional design used for data marts only
Dimensional Data Warehouse	Ralph Kimball	• Enterprise data warehouse • Bus architecture • Architected data marts • Virtual data marts	• *Dimensional data warehouse* is an integrated repository of atomic data • It may be accessed directly • Subject areas within the dimensional data warehouse • Data marts not required to be separate databases	All data is organised dimensionally
Stand-Alone Data Mart	No takers, yet common	• Data mart • Silo • Stovepipe • Island	• Subject area implementation without an enterprise context	May employ dimensional design

Operational Data Store

- Contains current or near current integrated data
- Subject oriented
- Limited amount of historical data
- Volatile
- Speed of data updates varies from seconds to a day
- Quick updating limits transformation possibilities
- Comes in different types with different levels of integration and quality

Data Warehouse Architecture Exercise

- You are asked by your company to propose a data warehouse architecture:
  - The director for a company wide solution
  - The manager of a department to give him specific information
  - The Operations Manager to help him to manage his operation
- Propose an architecture and explain your choice

Dimensional Modelling

Purpose of Analytic Databases

- Operational systems support the execution of business processes
- Analytic systems support the evaluation of processes
- Both systems have contrasting usage profiles
- Different principles guide their design
- Interaction with an analytic system takes place exclusively through queries that retrieve data
- These queries can involve large numbers of transactions
- It supports the maintenance of historic data

The Star Schema

- Dimensional design for a relational database
- Contains dimension and fact tables
- Dimension tables contain context for facts
- Dimensions are used to specify how facts will be rolled up
- Dimension values may be used to filter reports
- Dimension tables are not in third normal form

The Star Schema

- Each dimension table is given a surrogate key, typically an integer
- The dimension table key column name usually have the same suffix, like _key
- The dimension tables also contain columns that uniquely identify something in an operational system, like customer_id, salesperson_id, product_code. These are called natural keys
- By having separate surrogate keys and natural keys, you can track changes for dimension values
- Fact tables contain the facts and surrogate keys to the related dimension tables
- Often a fact row can be uniquely identified by these foreign keys, but not always
- The level of detail of the fact table is called the grain
- The information in the fact tables is typically consumed in different levels of details, using aggregation

Main Guiding Design Principles

- These two design principles are at the core of dimensional modelling:
  - accuracy
  - performance
- Accuracy: is it possible that facts can be aggregated in a way that does not make sense? Is there a design alternative that can prevent this?
- Performance: dimensional designs are very good to providing a rapid response to a wide range of unanticipated questions

Dimension Table Features - Keys

- Each dimension table is assigned a surrogate key. It is created especially for the data warehouse or data mart
- Surrogate keys are usually integers, generated and managed as part of the ETL process that loads the star schema
- One or more natural keys will also be present in most dimension tables
- The natural keys are identifiers carried over from source systems
- They identify a corresponding entity in the source system
- The values in natural keys may have meaning to users of the data warehouse
- Even without significant meaning, the presence is needed for the ETL that load fact tables

Dimension Table Features - Rich set of dimensions

- Dimensions can be added to queries in different combinations to answer a wide variety of questions
- The larger the set of dimension attributes, the more ways that facts can be analysed
- Dimension tables with a large number of attributes can be thought of as wide
- Commonly used combinations of attributes may be stored
- Codes may be supplemented with corresponding description values
- Flags are translated from boolean values into descriptive text
- Multi-part field are both preserved and broken down into constituent pieces
- Consider numeric attributes that can serve as dimensions

Dimension Table Features - Common Combinations

- In operational systems, it is common practice data elements down to constituent parts whenever possible
- In the dimensional design, common combinations of these elements are stored as well. Uses:
  - Increases query performance
  - Sort reports
  - Order data
- Example:
  - First name, middle initial, last name
  - Store also full name and Last-name-first format
  - Database administrators can index these columns for efficient query performance

Dimension Table Features - Codes and Flags

- In operational systems it is common to describe values in a domain using codes
- Both the codes and description may be useful dimensions
- Store both in your dimension table so that users can filter, access and organise in whatever way they see fit
- Flags can be stored in source systems in different ways; boolean data type, integer with value 0 or 1, character with "Y" or "N" or two values indicating "True" or "False"
- In a dimensional design, store the descriptive value of the flag options. These are far more useful than 0/1 or Y/N and much clearer when defining a query filter

Grouping Dimensions

- Dimension attributes are grouped into tables that represent major categories of reference.
- Junk dimensions collect miscellaneous attributes that do not share a natural affinity.
- When principles of normalisation are applied to a dimension table, the result is called a snowflake
- Snowflakes may be useful in the presence of specific software tools. Dimensional design fully embraces redundant storage of information (= no snowflakes)

Dimension Table Example

DIM_product
product_key
product_code
product_name
product_group
brand
size
colour
cost_price

Dimension Table Features - Benefits of Redundancy

- The storage of redundant data element specific in dimensional modelling have three advantages in an analytic environment:
  - performance
  - usability
  - consistency
- Precomputing and storing extra columns reduces the burden on the DBMS are query time, optimise performance with indexes and other techniques
- The redundant information makes it also easier for users to interact with the analytic database
- Explicit storage of all dimensions guarantees they are consistent, regardless of the application being used.

Degenerate dimensions

- Sometimes some dimensions associated with a business don't fit into a neat set of tables
- It may be appropriate to store one or more dimensions in the fact table. It is then called a degenerate dimension
- Although stored in the fact table, the column is still considered a dimension
- Consider if the attribute is really a degenerate dimension. Often such dimensions are better placed in junk dimensions.
- Transaction identifiers are commonly used as degenerate dimensions

Degenerate Dimension Example

FCT_order_line
order_date_key
customer_key
product_key
order_number
order_line
quantity_ordered
unit_price
discount_given

Slowly Changing Dimensions

- Information in a dimension table may change in the operational source over time, through correction of errors or updates.
- Because the dimension tables have surrogate keys as the primary key , it can handle changes different from the source systems
- How changes in source data are represented in dimension tables is referred to as slowly changing dimensions

Slowly Changing Dimensions - Type 1

- When the source of a dimension value changes, and it is not necessary to preserve its history in the star schema, type 1 is used
- The dimension (attribute) is simply overwritten with the new value
- The star carries no hint that the column ever contained a different value
- Any associated facts from before the change have their historic context altered
- Type 1 typically used for dimensions where a change is usually because of an error that is corrected (like birth date for a person)

Slowly Changing Dimensions - Type 2

- Type 2 preserves the history of facts:
  - Facts that describe events before the change are associated with the old value
  - Facts that describe events after the change are associated with the new value
- With type 2, a new row is inserted in the dimension table when there is a change in the source data
- This creates the effect of "versions" of a single dimension value in the dimension table
- These versions have the same natural key, but a different surrogate key value
- You can add a "current" flag to indicate the current row of a given natural key value
- To know when a version of a dimension row was valid, a date stamp is added

Choosing and Implementing Response Types

- A single dimension may have a type 1 response to some changes and type 2 response to other changes
- Most of the time a type 2 response is most appropriate
- There are situations in which the change of a source element may result in either type of response. When the source system records the reason for a change, you may choose to treat a change as type 1 in the case of an "error correction" or type 2 otherwise
- When a dimension contains multiple response types, ETL developers must factor in a variety of possible situations

Grouping Dimensions into Tables

- A dimensional model does not expose every relationship between attributes as a join
- Contextual relationships tend to pass through fact tables
- Natural affinities are represented by putting attributes in the same dimension table
- Dimensions are entities that can be related in multiple contexts (in different stars)
- Dimensions are grouped into tables based on natural affinity

Breaking Up Large Dimensions

- It is not uncommon for large dimensions to contain well over 100 attributes
- A dimension table may become so wide that it may have an effect on the database, like allocation of space or block size
- Large dimensions can be a concern for ETL developers. With many type 2 attributes, updates can become a tremendous bottleneck
- You may solve this by splitting dimensions arbitrarily
- An overwhelmingly large dimension may also be a sign that there are two distinct dimensions. Put these in two tables
- You can relocate free-form text fields to an outrigger

Dimension Roles and Aliasing

- Measurement of a business process can involve more than one instance of a dimension
- These roles are represented in a fact table by multiple foreign key references to the same dimension table
- This is very common to happen with the date dimension

Avoiding the NULL

- NULL can fail in WHERE clauses that lack a condition specifically for the NULL
- Never allow the storage of NULL in dimension columns. Instead, choose a value that will be used when data is not available (e.g. "Unknown")
- When a fact can't be associated with a row in a dimension table, we will use a special row in the dimension table
- You may have special rows for different situations, like invalid data or late-arriving data

Fact Table Features

- The fact table is the engine for business process measurement
- Where dimension tables are wide, fact tables are deep. They contain many more rows than dimension tables
- They contain foreign keys to the dimension tables, usually integers
- The facts themselves are usually integers or floating point decimal numbers
- The fact table should contain every fact relevant to the process it describes, even if some of the facts can be derived from others
- Some facts are nonadditive, like percentages or account balances

Fact Tables and Business Processes

- Dimensional models describe how people measure their world
- To be studied individually, each process should have its own fact table
- To determine if facts belong to one process, ask:
  - Do these facts occur simultaneously?
  - Are these facts at the same level of detail (or grain)?
- Multiple-process fact tables can be useful when *comparing* processes

Facts That Have Different Timing

- Events may share the same dimensions and seem related, but take place at different times. Then they are different processes and should have separate fact tables
- When a fact table for example can contain shipments and/or orders, the "and/or" in the statement of grain is usually a sign of problems to come
- Querying on such a table may get unexpected result rows that will confuse users
- Working around poor schema design may end up in an example of boiling the frog

Facts That Have Different Timing - Example

day_key	customer_key	product_key	quantity_ordered	quantity_shipped
123	777	111	100	0
123	777	222	200	0
123	777	333	50	0
456	777	111	0	100
456	777	222	0	75
789	777	222	0	125

These zeros will cause trouble

Facts That Have Different Timing - Example

Shipment Report - January 2008 - Customer 777

Product	Quantity shipped
Product 111	100
Product 222	200
Product 333	0

Page 1 of 1

A zero appears because there was an order

Facts That Have Different Grain

- When two or more facts describe events with different grain, they describe different processes
- Different grain can be caused by a different number of related dimensions or different level of hierarchy in a dimension (e.g. months versus days)

Fact Table Types

- The transaction fact table tracks individual activities or events that define a business process
- The snapshot fact table periodically samples status measurements such as balances or levels
- The accumulating snapshot table is used to track the progress of an individual item through a series of steps

Transaction Fact Tables

- Examples:
  - Booking of an order
  - Shipment of a product
  - Payment on a policy
- Each individual row describes the occurrence of an event
- By storing facts and associated dimensional detail, they allow activities to be studied individually and in aggregate

Transaction Fact Table Grain

- May be defined by referencing an actual transaction identifier, such as an order line
- May be specified in purely dimensional terms, as in "orders by day, customer, product and salesperson"
- Sometimes the grain is already a summary instead of an individual transaction, for instance because detail is available elsewhere or because the transaction volume is too large
- Despite a clearly defined grain, also an optional relationship is possible. Then the dimension contains a special row to represent this missing relation, like "not applicable"

Transaction Fact Tables Are Sparse

- Rows are only recorded for activities that take place, not for every combination of dimension values
- For instance rows are only created for those days when there are orders, only those products that are ordered and customers that place the orders

Transaction Fact Tables Contain Additive Facts

- Most nonadditive measurements, like ratios, can and should be broken down into fully additive components
- This allows the granular data in the fact table to be aggregated to any desired level of detail
- If you can use the sum of each measurement in the fact table in an aggregation, the fact is additive
- Storing fully additive facts provide the most flexible analytic solution

Transaction Fact Table Example

FCT_order_line
order_date_key
customer_key
product_key
order_number
order_line
quantity_ordered
unit_price
discount_given

Snapshot Fact Tables

- Are used to describe the effect of a series of transactions. These effects are called status measurements
- Some status measurements cannot be described as the effect of a series of transactions, for example the water level in a reservoir, the oxygen level in the air
- The snapshot fact table samples the measurement in question at a predetermined interval
- A snapshot fact eliminates the need to aggregate a long chain of transaction history

Snapshot Fact Example

- To know the balance of a bank account it is possible to calculate this from the full transaction history
- Over time this may involve thousands of transactions per bank account
- The account balance may be used to compute interest fees for example

When Transaction Data Is Not Stored

- It is possible that transactions reach further back into the past than is recorded in the data warehouse. For example a bank account that has been active for 50 years
- The volume of transaction detail may be too large to store in the data warehouse. For example the quality of train tracks every 20 cm
- A measurement may be status-oriented. For example budgets, temperature readings, reservoir levels

Don't Store the Balance with Each Transaction

- The transaction fact table is sparse. When there is no activity on a certain day, the balance will not be recorded when stored with transactions
- When there is more than one transaction, there will be double-counting in queries

The Snapshot Model

- Snapshots are dense
- A snapshot model contains at least one fact that is semi-addiditive
- The grain of a snapshot must include the periodicity at which status will be sampled and a definition of what is being sampled
- The grain of a snapshot fact table is usually declared in dimensional terms (definition of what is being sampled)

Semi-Additivity

- A semi-additive fact cannot be summed meaningfully across the time (date) dimension
- The fact can be additive across other dimensions
- The semi-additive fact can be summarised across periods in other ways, like minimum, maximum and average
- Some status measurements are not additive at all. For example water level or ambient temperature

Snapshot Fact Table Example

FCT_bank_balance
period_key
bank_account_key
branch_key
account_balance

Pairing Transaction and Snapshot Designs

- Many processes can be modelled both in a transaction and a snapshot fact
- When a design will include both a transaction fact table and a periodic snapshot, the snapshot can and should be designed to use the transaction fact table as a source
- This eliminates duplicative ETL processing of the source data
- It ensures that dimensional data will be identified and loaded consistently

Accumulating Fact Tables

- Focuses on time between events in a process
- The grain is a unit that goes through the business process, like a loan application
- The fact table will have exactly one row for each unit
- It will have multiple keys to the Date dimension for completion of each stage of the process
- Each row has a group of facts that measure the number of days spent on each stage

Accumulating Fact Tables

- The active rows are updated regularly
- Fact for the duration of the active step is incremented at each load
- Each time a stage is completed, the appropriate end date key is set
- When the design for a business process includes both a transactional star and an accumulating snapshot, the accumulating snapshot should use the transaction star as its source

Dimensional Modelling Exercise

- You are approached by one department of your company to create a data mart for one of their processes:
  - Accounting - bookkeeping
  - Sales - product sales
  - Human Resources - employees
  - Specific to company:
    - Production
    - Client product development
    - Customer activity (telecom)

Querying Dimensional Models

Using a Star Schema

- Most queries against a star schema follow a consistent pattern:
  - One or more facts are requested, along with the dimensional attributes that provide the desired context
  - The facts will be summarised in accordance with the dimensions present in the query
  - Dimension values are used to limit the scope of the query (filter)
- The star schema can be used in this way with any combination of facts and dimensions (in the star)
- Note that the ability to report facts is primarily limited by the level of detail at which they are stored.
- Various aggregations are sum, average, count

Typical Star Schema Query Example

```
SELECT store_location, month_name, SUM(sales_price) AS
total_sales, SUM(discount) AS total_discount
FROM fact_sales fs
JOIN dim_date dd
ON dd.date_key = fs.date_key
JOIN dim_sales_people dp
ON dp.sales_people_key = fs.sales_people_key
WHERE year = 2015
AND country = 'Belarus'
GROUP BY store_location, month_name
ORDER BY month_number
```

see: aggregation, relate fact table to dimension tables, filters, order

Typical Star Schema Query Example Alternative

```
SELECT store_location, month_name, SUM(sales_price) AS
total_sales, SUM(discount) AS total_discount
FROM fact_sales fs, dim_date dd, dim_sales_people dp
WHERE dd.date_key = fs.date_key
AND dp.sales_people_key = fs.sales_people_key
AND country = 'Belarus'
AND year = 2015
GROUP BY store_location, month_name
ORDER BY month_number
```

Analysing Facts From More Than One Fact Table

- When comparing facts from different fact tables, it is important to collect them from separate SELECT clauses
- When you use a single SELECT, there is risk of double counting, or worse
- The two-step process used is called *drilling across*, stepping from one star to another

Drill-across Procedure

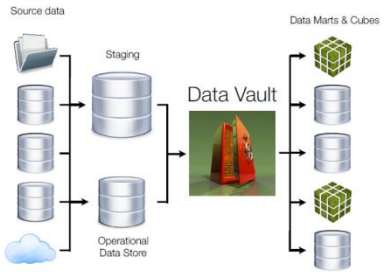
- Phase 1: retrieve facts from each fact table, applying appropriate filters, outputted in desired level of dimensional detail
- Phase 2: merge the intermediate results together
- This process can be done with any amount of fact tables
- This can also be done across different databases, as long as the dimensions involved have the same structure and content

How to Query Semi-Additive Facts

- When summing the semi-additive fact, the query must be constrained (filtered) by a unique row in the nonadditive dimension, or grouped by rows in the nonadditive dimension
- Consider the grain of the snapshot fact table to see if the SQL AVG function can be used

Data Vault

Data Vault in a Data Warehouse Architecture



Data Vault Fundamentals - Hub

- The Hub represents a core business concept as Customer, Vendor, Sale, Employee
- The hub table is formed around the Business Key of this concept
- A hub row is created the first time a specific business key is introduced to the Enterprise Data Warehouse
- The hub contains no descriptive information and no foreign keys
- The hub contains only the business key, a data warehouse ID, a load date-timestamp and a record source

Data Vault Fundamentals - Hub

H_customer
h_customer_sid
h_customer_code
h_customer_ldts
h_customer_record_source

Data Vault Fundamentals - Link

- A Link represents a natural business relationship between two or more business keys
- Just like the hub, it contains no descriptive information
- A link row is created the first time a unique association between business keys is introduced to the Enterprise Data Warehouse
- The link consists of the data warehouse IDs from the hubs that it is relating, with a data warehouse ID, a load date-timestamp and a record source

Data Vault Fundamentals - Link

L_customer_product_sale
lnk_cps_sid
lnk_customer_sid
h_product_sid
h_sale_sid
lnk_cps_ldts
lnk_cps_record_source

Data Vault Fundamentals - Satellite

- The Satellite contains the descriptive information or context for a business key
- There can be several satellites to describe a single business key (hub) or association of keys (link)
- A satellite can describe only one key (hub or link)
- The satellite is connected to a hub or link with the data warehouse ID of the hub or link
- The key of a satellite row is the hub or link key and the date-timestamp
- The satellite is the only construct that manages date warehouse history using various rows with date-timestamps to record the validity of each row
- A satellite has no foreign key constraints

Data Vault Fundamentals - Satellite

S_customer
h_customer_sid
s_customer_ldts
s_customer_ledts
customer_name
customer_address
s_customer_record_source

Choosing Satellites

- There are different reasons to put attributes or context in various satellites:
  - subject area
  - rate of change (do values change often or seldom)
  - source system (and arrival time of data)

Modelling With The Data Vault

- Identify business concepts
- Establish the enterprise wide business keys for hubs
- Model the hubs
- Identify natural business relationships
- Analyse relationships Unit of Work (relationships formed from a business perspective)
- Model the links
- Gather context attributes to keys
- Establish criteria and design satellites
- Model the satellites

Data Vault Modelling Challenges - Business Keys

- A business key is a unique identifier according to a business person
- Some business concepts may lack a visible identifier

Data Vault Load Order

- First load the hubs, so new keys are appointed to new rows in the hubs
- Secondly load the links, so new keys are appointed to new rows and the correct hub data vault keys can be assigned to each row
- Lastly load the satellites, so the correct hub or link data vault keys can be assigned to each row

Data Vault 1 or 2 - ID

- The original Data Vault uses a meaningless sequence (integer) per hub or link as an ID
- The new Data Vault uses a hash key derived from the business key
- The hash key has the advantage that parallel loading of hubs, links and satellites is possible
- The hash key ID can cause key collisions (identical keys), although the chance of this is tiny

Data Vault Advantages

- Uses mainly fast inserts into the database instead of slower updates
- Restarting a load again after an error can be done safely
- Using many-to-many relationships by default means no rework when the relationship type changes
- Traceability with the load date-timestamp and record source columns
- Use of various satellites offers flexibility and means no rework when new attributes are added or source systems change
- System of 'facts' as there is (almost) no application of business rules, cleansing or other transformations

Data Vault Considerations

- Data Vault is bad for querying, it is no substitute for data marts
- The amount of tables is higher due to the separation in hubs, links and satellites

Data Vault Exercise

- Create a Data Vault model that fits the dimensional model you created earlier
- Do it step by step:
  - Hubs
  - Links
  - Satellites
- Present and discuss results after each step

MIR-3179 CAN BE REPRESS APOPTOSIS IN NSCLC

H. AKCA, PhD, Professor
Pamukkale University, Medical Faculty, Department of Medical Biology

Ş. AKGUN
Research Asistant, Pamukkale University, School of Medicine, Medical Biology Department

Pamukkale University, School of Medicine, Medical Biology Depertmant, Denizli, Turkey

Lung Cancer is the most common cancer type resulting in death of the course of disease world-wide. Approximately 85% of lung cancer cases are Non-Small Cell Lung Cancer (NSCLC). MicroRNAs (miRNAs) have a key role in post-transcriptional regulation binding to the 3'UTRs of mRNAs. Recent studies have shown that miRNAs may be potential target at NSCLC theraphy. We aimed to find NSCLC related miRNAs at our study. We selected some miRNAs that might potentially be related with NSCLC according to our bioinformatic analysis. We investigated that expression level of miR-3179 on NSCLC was higher than breast cancer and normal endothelial cell lines. So, we generated microarray analysis for finding out miR-3179 target mRNAs and role at biological proces of that. Our microarray pathway results showed that miR-3179 is highly related with apoptosis and Wnt signaling pathway.

C3435T POLYMORPHISM OF MDR1 GENE EFFECT OF SURVIVAL ON NON-SMALL CELL LUNG CANCER

N. KARAGENÇ, MD, PhD
Assistant Professor, Pamukkale University, Medical Faculty, Department of Medical Biology

E.R. KARAGUR
Research Assistant in Pamukkale University, Faculty of Medicine, Department of Medical Biology

O. TOKGUN
Research Assistant in Pamukkale University, Faculty of Medicine, Department of Medical Biology

A. DEMIRAY, PhD
Assistant Professor, Pamukkale University, Medical faculty, Dept of Medical Biology

Ş. AKGUN
Research Asistant, Pamukkale University, School of Medicine, Medical Biology Department

H. AKCA, PhD, Professor
Pamukkale University, Medical Faculty, Department of Medical Biology

Pamukkale University, School of Medicine, Medical Biology Depertmant, Denizli, Turkey
E-mail: nkaragenc@hotmail.com

Purpose: MDR1 gene which codes P-glycoprotein (PGP) is important factor chemotherapy resistance. The expression of P-glycoprotein (PGP) is higher TT genotype with patient than CC genotype with patient. In additionally TT genotype with patient has good progression chemotherapy treatment. We have evaluated whether the genotypic and allelic polymorphism has the effect of survival on MDR1 gene codon 3435.

*Material and Method:*DNA was isolated 79 patient with Non-small cell lung cancer which treatment chemotherapy by qiagen EZ1 kit. C3435T polymorphism was evaluated pyrosequencing and qRT-PCR.

*Results:*The average age was 59.8±7.8 years, 68.7% of patients are male, 31.2% are female. Genotype frequency of CC, CT and TT is detected as 35.4%, 51.6% and 12.9%, and allele frequency of T and C as 44.5% and 55.5%, respectively. All patients have been followed up for 2 years. The elapsed period until progression was calculated as 21±8.7months in CC genotype, 13±1.6 months in CT genotype and 12±3.5 months in TT genotype but no significant difference was found between them. General survival period was 28.9± 4.7 months and these value 52.5± 12.4 months for CC genotype, 20.5± 2.7 months for CT and TT genotype and Statistically significant(p=0.004).

*Conclusion:*Our results indicate that CC genotype was less progression and related with general survival. The results from our patient group correlate with the literature.

MOLECULAR SPECTRUM OF KRAS, NRAS AND BRAF MUTATIONS IN DENIZLI COLORECTAL CANCER PATIENTS

H. AKCA, PhD, Professor
Pamukkale University, Medical Faculty, Department of Medical Biology

E.R. KARAGUR
Research Assistant in Pamukkale University, Faculty of Medicine, Department of Medical Biology

A. DEMIRAY, PhD
Assistant Professor, Pamukkale University, Medical faculty, Dept of Medical Biology

Ş. AKGUN
Research Asistant, Pamukkale University, School of Medicine, Medical Biology Department

O. TOKGUN
Research Assistant in Pamukkale University, Faculty of Medicine, Department of Medical Biology

N. KARAGENÇ, MD Ph
Assistant Professor, Pamukkale University, Medical Faculty, Department of Medical Biology

Pamukkale University, School of Medicine, Medical Biology Depertmant, Denizli, Turkey
E-mail: nkaragenc@hotmail.com

Purpose: Mutations in genes such as KRAS, NRAS and BRAF have become an important part of colorectal carcinoma evaluation. The aim of this study was to screen for mutations in these genes in Turkey patients with colorectal cancer (CRC) and to explore their correlations with certain clinicopathological parameters.

Material and Method: We tested mutations in the KRAS (exons 12, 13 and 61), NRAS (exons 12, 13 and 61), and BRAF (codon 464 and 600) genes using polymerase chain reaction with biotinylated primers following pyrosequencing in a small portion of 136 Turkish CRC patients who has applied to Pamukkale University Hospital.

Results: The prevalence rates of KRAS, NRAS and BRAF mutations were 45%, 15% and 8%, respectively. Mutant KRAS was associated with the mucinous subtype and greater differentiation, while mutant BRAF was associated with right-sided tumors and poorer differentiation.

Conclusion: Our results revealed that correlation in the genetic profiles of KRAS, NRAS and BRAF at mutation hotspots in Turkish CRC patients and some of those mutations patterns were consisted with those patients from the far East countries.

DEVELOPED RLE ALGORITHM AND BITPLANE SLICING TO COMPRESS GRAYSCALE IMAGE

H. K. ALBAHADILY
Aspirant in telecommunication and computer network, BSUIR

V. TSVIATKOU, PhD
Head the Department of the Telecommunication Systems, BSUIR

Belarusian State University of Informatics and Radio Electronics, Republic of Belarus
E-mail: hasan@bsuir.by

Abstract. New suggested RLE compression algorithm to compress grayscale images with bitplane slicing technique to reduce the size of the encoded data by separating image into 8 binary layers, then use our modified RLE algorithm to compress the bitplanes. Our modified algorithm designed perfectly to compress bitplane. The proposed method achieved very good compression ratio especially with the MSB layer.

Introduction. Data files frequently contain the same character repeated many times in a row or column. The digitized signals can also have runs of the same value, indicating that the signal is not changing, also images and music [1]. The Image can be considered as a two dimensional array of pixel intensities or can be considered as a discrete representation of data possessing both spatial (layout) and intensity (color) information [2].

There is significant redundancy present in image signals. This redundancy is proportional to the amount of correlation among the image data samples [3].

The goal of image compression is to represent an image signal with the smallest possible number of bits, thereby speeding up transmission, minimizing storage requirements,reduces the cost of data transmissionand reduces the errors of transmission.

Our method is implemented using MATLAB2012 on WINDOWS7 Operating System.

Bitplane slicing. The bitplane slicing is a fundamental technique of image processing in which the image is sliced into different planes(each layer contains sequences of only binary digits 0 or 1). It is ranges from plane1 which contains the least significant bit (LSB) to the last planeN which contains the most significant bit (MSB), where the number of layers depends on the bit depth of the image. The bit depth means how many bits need to represent the pixel's intensity. for example if the image is grayscale i.e. bitdepth is 8bit and it will be separated into 8 layers, or into 24 layers if the image is colored i.e. bitdepth is 24bit.

It is clear that the intensity value of each pixel can be represented by 8-bit binary vector (b8,b7, b6, b5, b4, b3, b2, b1) bk, where k is from 1 to 8 and each bk is either "0" or"1". In this case, an image may be considered as an overlay of eight bit-planes. Each bit-plane can be thought of as a two tone image and can be represented by a binary matrix [6][7].The formation of $bitplane_k$ is given by Equation below [4]:

$$BitPlane_K = Reminder\left\{\frac{1}{2} floor\left[\frac{1}{2^{K-1}}\ Image\ \right]\right\} \quad (1)$$

The bitplane decomposition is very useful for image compression. It allows some bi-level compression [5], i.e. it is used in some ways to compress images based on the idea of splitting the image

into layers of binary values then either omit layers which are not highly effect on the image quality or by using the idea of similarity of elements in the bit plane which would be appears highly in the MSB layers, and by this way a long runs of similar values would result in very good compression rates [8][9]. Thus The RLE may be advantageously applied because the long runs in the bit planes which is the backbone of RLE [10]. This technique is very useful even if there are no repeated runs in the pixels, and by using bitplane slicing we will find some kind of repetition especially with last layer which contains MSB and achieving the highest compression ratio because it is contain frequently repeated runs.

The modified RLE algorithm (I3BN). The RLE method counts the values and their runs or repeated time as pairs of value and run (I,N), where I is the vector of values and N is the vector of repeats.

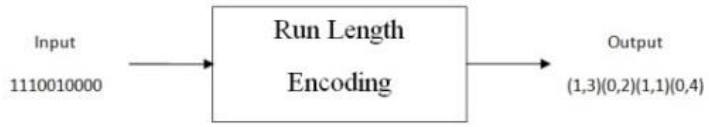


Fig. 1. Illustration of RLE for a binary input sequence

The modified RLE algorithm I3BN using the same idea of RLE which countsrepeated runs but instead of sending the values and runs as pairs we will send only the repeated runs by sending one bit or two or three followed by the number of repeated runs [11].

The algorithm I3BN using three symbol b1, b2 and b3, which takes 1 if the value I repeated and 0 if absent. The structure of coded data for algorithm I3BN can be represented by the following diagram:

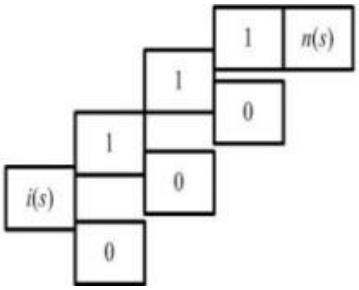


Fig. 2. Coded data structure of I3BN

So if the run repeated one time we will send the binary series 0, if the run repeated two times we will send the binary series 10, if the run repeated three times we will send 110 and if the run repeated more than three times we will send the sequence 111 followed by subtraction four from the run, for example if the value repeated 8 times, we will send 111(4), and we need to represent the run in binary and reserve number of bits to represent the new subtracted runby finding the binary logarithm which will be 3 bits so we will send 111 (100).

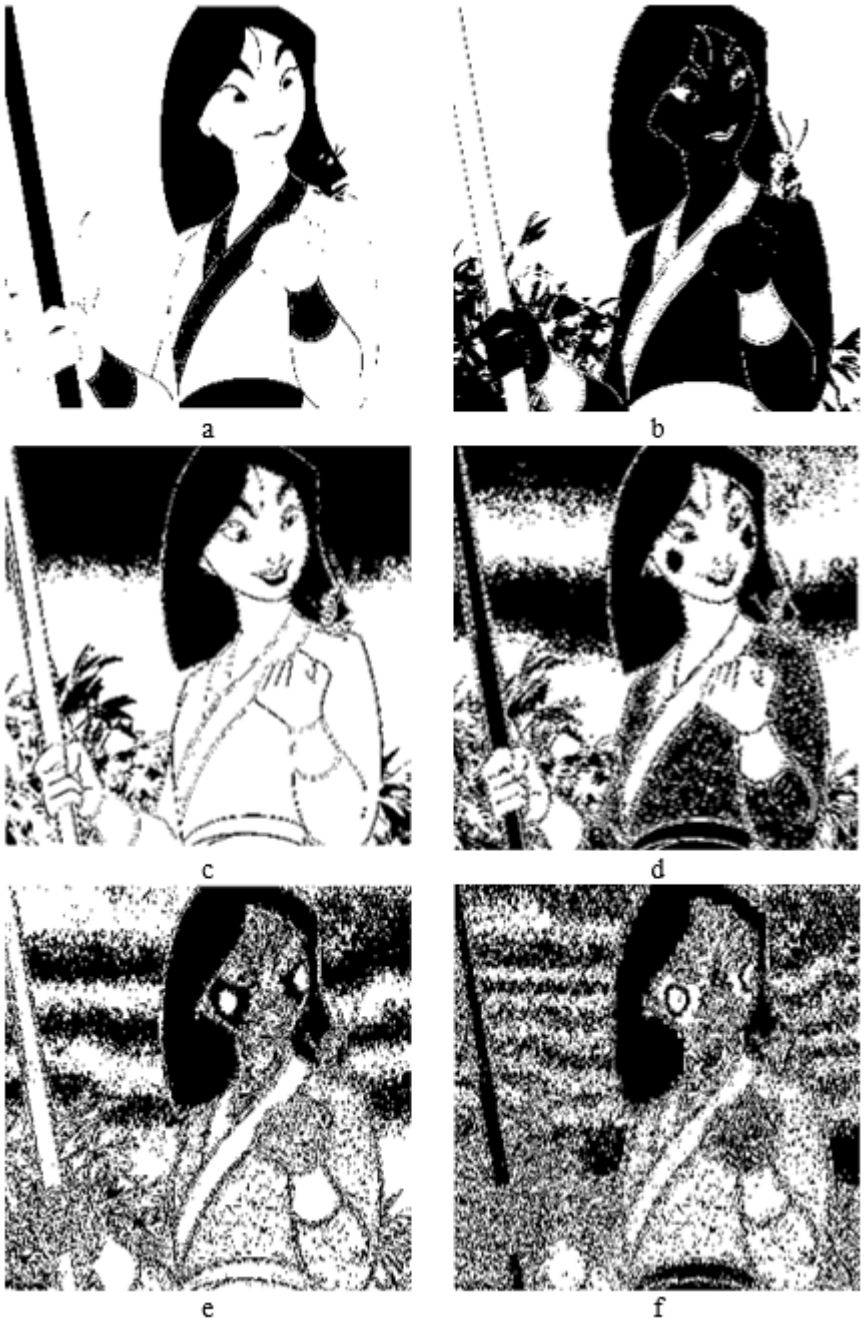
The code Size R_{I3BN} (bit) ofalgorithm I3BN defined by the expression:

$$R_{I3BN} = S(BD_I + 1) + \sum_{s=0}^{S-1} b1(s) + \sum_{s=0}^{S-1} b2(s) + BD_N \sum_{s=0}^{S-1} b3(s) \quad (2)$$

Where BD_I is the bit depth of the image and BD_N is the bitdepth of the maximum repeat.

The modified algorithm I3BN will work perfectly with bitplane by increasing the number of runs and decreasing the number of bits to represents runs.

Experiment implementation of bitplane slicing. First step is converting the images into bit-planes(each layer will be binary image contains 0,1 only)



a - bitplane8; b - bitplane7; c - bitplane6; d - bitplane5; e - bitplane4; f - bitplane3

Fig. 3. The bitplanes for test image

It is possible to remove some information from an image without any apparent change in its visual appearance because the first three bits does not contribute so much information in image formation.

The image can be stored with the information provided by bit4 to bit8 only. Thus number of bits per pixel can be reduced to 5 which save more storage space [2][4].

Fig. 4. The test image with their 5bit reduced image

We can see that there is no big change in the visual appearance of the images because the 3 LSB does not contribute big value.

Results and disscution. First part of experiment is compressing the original test images without bitplanes, which achieved not high compression ratio as we can see from the table below:

Table 1. Compression ratio without bitplane

Image	Img1	Img2	Img3
CR	1,082	1,053	0,928

From results above we can see that the proposed algorithm provides compression ratio up to 1,082-0,928 times for the original images without using bitplane technique.

The second part is splitting the images into 8 bitplanes and implementing the modified RLE algorithm I3BN on each plane, and we got the result in the table below:

Table 2. Compression ratio with bitplane

Image	Layer8	Layer7	Layer6	Layer5	Layer4	Layer3	Layer2	Layer1
Img1	3,134	2,294	1,669	1,039	0,863	0,765	0,681	0,638
Img2	3,551	3,221	2,354	1,853	1,213	0,744	0,638	0,608
Img3	2,206	1,385	1,116	0,913	0,752	0,638	0,616	0,605

From results above we can see that the proposed algorithm provides compression ratio up to 3,551-0,605 times for the MSB plane of the images.

The third part is implementing the modified algorithm on the images after reducing the three LSB and we got the results below:

Table 3. Compression ratio for reduced images

Image	Img1	Img2	Img3
CR	2,028	2,970	1,572

We can see that the modified algorithm has achieved very good ratio of compression up to 2,970-1,572 times without big change in the image visual details.

Conclusion and future work. The Results showing that the modified algorithm provides compression ratio for the original images without bitplane up to 1,082-0,928 times while the compression ratio was up to 3,551-0,605 times for bitplanes of images. The modified algorithms provide encoded image size reduction up to 2,970-1,572 times when removing the first three LSB bits.

For the future work it will be good to use the proposed algorithm with other compression methods like JPEG to get a high compression ratio in the lossy compression methods with keeping the image visual details as high as we can.

References

[1]. S. Smith The Scientist and Engineer's Guide to Digital Signal Processing 2nd Edition, California Technical Publishing 1999.

[2]. C. Solomon, T. Breckon Fundamentals of Digital Image Processing, A Practical Approach with Examples in Matlab , John Wiley Ltd. 2011.

[3]. A. Bovik The Essential Guide to image Processing, Elsevier Inc. USA 2009.

[4]. S. Halder [et al] A Low Space Bit-Plane Slicing Based Image Storage Method using Extended JPEG Format, International Journal of Emerging Technology and Advanced Engineering Vol 2, Issue 4, April 2012.

[5]. M. Luís O. Matos Lossy-to-Lossless Compression of Biomedical Images Based on Image Decomposition, Applications of Digital Signal Processing through Practical Approach, InTech.

[6]. R. Gonzalez, R. Woods Digital Image Processing 2nd ed., Prentice Hall, Inc., New Jercy 2008.

[7]. K. Ting, D. Bong, Y. Wang Performance Analysis of Single and Combined Bit-Planes Feature Extraction for Recognition in Face Expression Database, International Conference on Computer and Communication Engineering 2008, 2008 Kuala Lumpur, Malaysia.

[8]. M. Rangaraj Biomedical Image Analysis, University of Calgary, Canada, CRC Press LLC 2005.

[9]. M. Alasdair An Introduction to Digital Image Processing with Matlab, Victoria University of Technology, 2004.

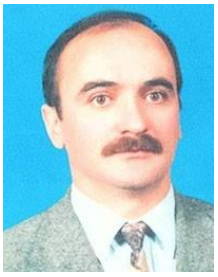
[10]. A. BOVIK Hand Book Of Image And video Processing, Academic Press Austin USA 2000.

[11]. H. Albahadily [et al] New Modified RLE Algorithms to Compress Grayscale Images with Lossy and Lossless Compression, International Journal of Advanced Computer Science and Applications, Vol. 7, No. 7, 2016.

THREE DIMENSIONAL (3D) IMAGING METHODS ON PATIENTS WITH ISCHEMIA POINTING OF FUNCTIONAL REGION IN BRAIN BY CLINICAL SIGNS

M.B. ÖZDEMIR, MD PhD, Professor[1]
Pamukkale University, Medical Faculty, Department of Anatomi

N. KARAGENÇ, MD Ph
Assistant Professor, Pamukkale University, Medical Faculty, Department of Medical Biology



A. AYDIN, PhD, Professor[2]
Pamukkale University FAculty of Engineering, Department of Geophysics

[1]Pamukkale University, Medical Faculty, Anatomy Dept, Denizli, Türkiye
[2]Pamukkale University, Engineering Faculty, Denizli, Türkiye
E-mail: nkaragenc@hotmail.com

Key words: Brain, ischemia, MRI, CT, clinics, neurology, symptoms, 3D, computational neuroscience.

Objective: The relationship between neurological findings in patients with ischemic damaged brain regions shown in more postmortem studies. However, 3-dimensional (3D), the case has not been evaluated in vivo. The purpose of this study, magnetic resonance obtained from the patient (MR) and computed tomography (CT) three sectional ischemic damaged brain regions on images dimensional (3D) is by making investigate and to correlate them with the clinical findings.

Methods: 105 patients for this purpose (53 males, 52 females) were examined images in 3D computer-assisted programs and clinical findings were correlated with infarct scale. Level of consciousness, orientation, limb motor activation, the facial motor activity, eye movements, visual fields, limb ataxia, sensory conditions, neglect articulation and language were evaluated.

Results: In the clinical signs of ischemia under the influence of men and women were found to be different. Unlike known, damage was observed in patients with ischemic area of the same lead to different clinical manifestations. Infarct size was found to be an important factor in the emergence of clinical scheme.

Conclusion: A plurality of functional regions of the brain has been identified to date. Present study was performed to correlate the clinical evaluation in 3D for the first time. The results of computer-aided neuroscience (compututional neuroscience) can be the source terms. As a basic and clinic could pave the way for new studies

CLOUD COMPUTING IN EDUCATION

T. TANYERİ
Assistant Professor, Pamukkale University, Faculty of Education, Dept of Computer and teaching Technologies in Education

H. KIRAN, Professor
Professor Pamukkale University, Faculty of Education, Dept of Primary Education

Pamukkale University, Faculty of Education, Denizli, Türkiye
E-mail: ttanyeri@pau.edu.tr, hkiran@pau.edu.tr

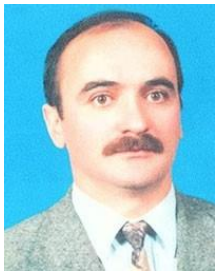
Key words: cloud computing, educational technology, innovation

The developments in the digital world show us that even the dreams established in the past days have exceeded their limits. The internet, which only transmits the limited data of military institutions and major trading companies, is one of the essential infrastructure elements of every sector today. At this point, it is seen that the internet is not only the basic necessity of the institutions or organizations but also the basic necessity of the person. For this reason, it is unlikely that the education sector will fall behind internet technologies. Cloud computing technology is one of the internet technologies that find solutions to many problems today. Cloud computing can be defined as a kind of software and hardware solution that is independent of time and place for people or institutions that need IT infrastructure. In other words Cloud computing is mainly is a computing service consisting of services provided by distant servers whose maintenances and hosting is carried out by others. But behind this simple definition is a complex technology. The use of cloud technology has been a game changer, as it is in all other sectors. With the use of Cloud Computing, infrastructure installation, management and updating etc. are performed by the professionals. On this count, the instructor, the manager and the students have the opportunity to allocate more resources to the teaching activities which are the main tasks. With cloud technology, the obligation to invest in continuous infrastructure upgrades in education and the requirement to acquire new software licenses ceases to exist. It is possible to have quick access to information resources. In this context, it is thought that the effect on future learning environments of Cloud Computing will increase in the future because it offers higher quality services by lowering costs.

MORPHOLOGIC-VOLUMETRIC ALTERATIONS IN BRAIN STRUCTURES RELATED WITH PSYCHOTIC DISORDERS INCLUDING SCHIZOPHRENIA, SCHIZOAFFECTIVE DISORDER AND PSYCHOTIC BIPOLAR DISORDER IN THE SAME STUDY FROM MRI

M.B. ÖZDEMIR, MD PhD, Professor[1]
Pamukkale University, Medical Faculty, Department of Anatomi

N. KARAGENÇ, MD Ph
Assistant Professor, Pamukkale University, Medical Faculty, Department of Medical Biology



A. AYDIN, PhD, Professor[2]
Pamukkale University FAculty of Engineering, Department of Geophysics

[1]Pamukkale University, Medical Faculty, Anatomy Dept, Denizli, Türkiye
[2]Pamukkale University, Engineering Faculty, Denizli, Türkiye
E-mail: nkaragenc@hotmail.com

Key words: anatomy, brain, morphometry, psychotic disorder, schizophrenia schizoaffecti disorder, psychotic bipolar disorder.

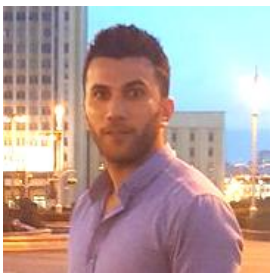
Objective: It has already been investigated that schizophrenia (SZ), schizoaffective disorder (SZA) and psychotic bipolar disorder (BD) cause volumetric alteration on brain structures previously. However, these are in separated studies from each other and have some contradictions in findings. The aim of present study is to estimate volume of the brain structures in the same study in order to improve understanding of morphologic abnormalities in underlined psychotic disorder.

Methods: For this purpose, brain structures from MR images of the 174 cases with psychotic disorder (58 female and 116 male) were compared with 186 healthy controls (67 female and 119 male). 16 right, 16 left and 11common, tottally 43 structures that might be related with psychotic disorders was evaluated.

Results: There was a volume decreasing in almost all structures in patient with SZ. But, ventricles volume increased in patient with all SZ, BD and SZA. Most of the alterations was correlated with Positive and Negative Syndrome Scale (PANNS).

Conclusion: Then we concluded that the effect of the psychotic disorders were definetely different on sexes. Volumetric alterations were descriptive mostly for patients with SZ. BD and SZ might overlap in clinical and biological features but they demonstrated significantly different alterations morphologically. PANNS was more correlated with SZ, especially with SZA than BD via morphometry. Morphemetric abnormality was less in BP than SZ and SZA. These findings indicate the availability of anatomical markers in the diagnosis and treatment of psychotic patients.

SELECTION TEXTURE REGIONS ON THE IMAGE BASED ON CLASSIFICATION ASSESSMENT DENSITY OF CONTOUR ELEMENTS



H.M. ALZAKKI
Aspirant in telecommunication and computer network, BSUIR

V. TSVIATKOU, PhD
Head the Department of the Telecommunication Systems, BSUIR

The Belarusian State University of Informatics and Radio Electronics, Republic of Belarus
E-mail: Haidermakki300@yahoo.com

Abstract. A method for texture images segmentation based on selection texture regions on the image based on classification assessment density of contour elements. The goal of the method find the contouring of the image, determining the position of contour elements in the image and classify it for different types(points, lines, and shapes) close the region which had same type of contour type into binary regions objects. The result will be representing in binary matrix.

Introduction. Texture segmentation and contour analysis provide an important information for machine vision tasks such as scene classification, surface orientation, and shape determination and so on. Contour analysis (e.g. edge detection) may be adequate for untextured images, but in a textured region it results in a meaningless tangled web of contours. For example, the detection of the edge will return to the region in the beans as shown in Fig. 1. The all old solutions problem in edge detection went to use a high threshold so as to minimize the number of edges which can found in the texture area. This is obviously a non-solution—such an approach means that low-contrast extended contours will be missed as well [4].

Fig. 1. Use Prewitt and Roberts filters for the image using different thresholds:
(a) low threshold level, (b) high threshold level.

In this paper; we proposed a method is to develop iterative algorithm selection texture regions on the image based on classification assessment density of contour elements [2, 3]. We are chosen

texture images from Brodatz textures with different and the method texture image segmentation based on classification of contour elements and logical addition of classes [5] are used to classify the image to different types like (point, line, cell, spot) and find the central pixel for each class and use the proposed method to find the homogenous (texture) regions in the image as shown in Fig. 2.

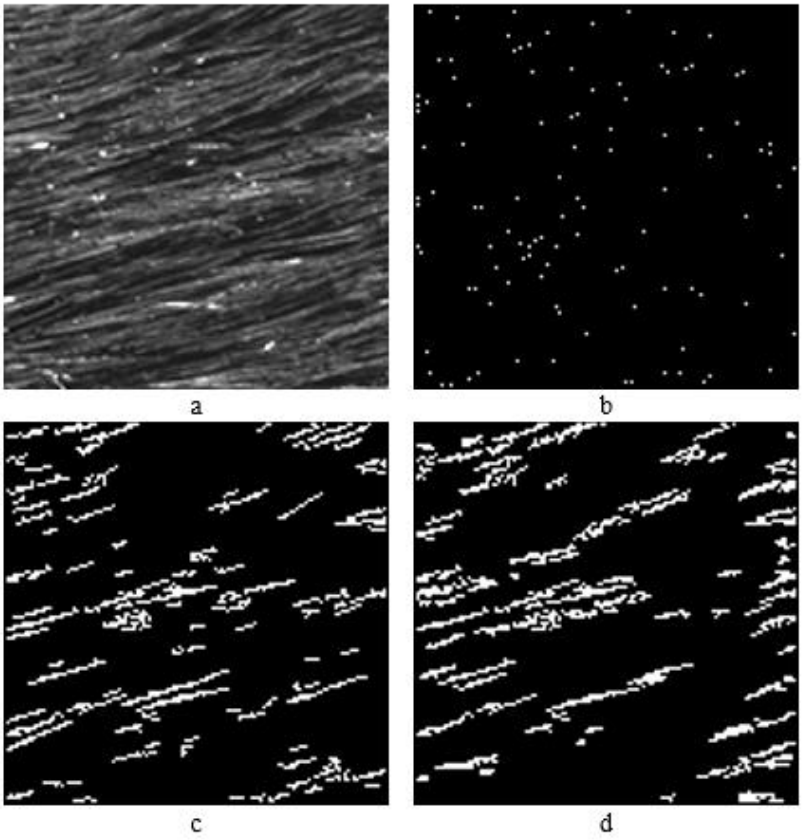


Fig. 2. (a) texture image, (b) point in image (a), (c) line in image (a), (d) spot in image (a)

The algorithm of selection texture regions on the image based on classification assessment density of contour elements. The algorithm selection texture regions on the image based on classification assessment density of contour elements the input image for the algorithm is a binary image

$$B = \left\| b(y,x) \right\|_{\left(y=\overline{0,Y-1},\, x=\overline{0,X-1}\right)} \tag{1}$$

of isolated homogeneous regions. The maximum allowable number of iterations $\hat{N}_C$ and the segment area $\hat{S}_S$, $\hat{N}_S$ the segments number where Y, X the size of the input image vertically and horizontally, $b(y,x)$ - A pixel of the input image, takes the value 1- for homogeneous segments and 0 - for the zone. The algorithm allows combining isolated regions in the larger region and then combines them until will be provided with the specified output conditions of the algorithm ($\hat{N}_C$ the maximum number of iterations, $\hat{S}_S$ segment area, $\hat{N}_S$ the number of segments).

The algorithm consists of the following steps:

1) Start the iteration

$n_C \leftarrow 0$

2) Start the iteration and combination of isolated homogeneous regions.

Segmentation of the homogeneous regions in the image. In this step the input image B is associated with a matrix

$$S_B = \left\| s_B(y,x) \right\|_{\left(y=\overline{0,Y-1},\, x=\overline{0,X-1}\right)} \tag{2}$$

the value of each element

$$s_B(y,x) \in [0, N_S] \tag{3}$$

which indicates for segment number, which it belongs, where N_S -The number of segments. Elements $s_B(y,x)=0$ they belong to the zone. For the image segmentation in this step are used the method of region growing [1].

3) Define the area for the segments. For all segmentsFormed by the pixels of the input image, the values which $b(y,x)=1$ segment $\mathrm{S}_S(n)$all pixels (the area) of the segments by using the equation

$$\left(b(y,x)=1\right) \Rightarrow \left(\mathrm{S}_S\left(s_B(y,x)\right) \leftarrow \mathrm{S}_S(n)+1\right) \tag{4}$$

when $x=\overline{0,X-1}$, $y=\overline{0,Y-1}$, where $n \in [1, N_S]$ – the number of the segment, when initialization $\mathrm{S}_S(n) \leftarrow 0$ when $n=\overline{1,N_S}$

4) Determine the maximum area of the segments by using the equation

$$\left(\mathrm{S}_S(n) > \mathrm{S}_{\max}\right) \Rightarrow \left(\mathrm{S}_{\max} \leftarrow \mathrm{S}_S(n)\right) \tag{5}$$

when $n=\overline{1,N_S}$ when initialization $\mathrm{S}_{\max} \leftarrow 0$

5) Define the size of the segments. For all segments formed by the pixels of the input image, the values which $b(y,x)=1$. Calculate the coordinates of the left $x_L(n)$, right $x_R(n)$, upper $y_H(n)$, lower $y_N(n)$ pixels by using the equation

$$\begin{aligned}\left(b(y,x)=1\right) \Rightarrow & \left(\left(x < x_L\left(s_B(y,x)\right)\right) \Rightarrow \left(x_L\left(s_B(y,x)\right) \leftarrow x\right)\right) \\ & \left(\left(x > x_R\left(s_B(y,x)\right)\right) \Rightarrow \left(x_R\left(s_B(y,x)\right) \leftarrow x\right)\right) \\ & \left(\left(y < y_H\left(s_B(y,x)\right)\right) \Rightarrow \left(y_H\left(s_B(y,x)\right) \leftarrow y\right)\right) \\ & \left(\left(y > y_N\left(s_B(y,x)\right)\right) \Rightarrow \left(y_N\left(s_B(y,x)\right) \leftarrow y\right)\right)\end{aligned} \tag{6}$$

when $x=\overline{0,X-1}$, $y=\overline{0,Y-1}$when initialization $x_L(n) \leftarrow X-1$, $x_R(n) \leftarrow 0$, $y_H(n) \leftarrow 0$ when $n=\overline{1,N_S}$

6) Calculate the coordinates for the centers for each segment. For all segments calculate the coordinates of their centers by using the equation

$$\begin{aligned} x_C(n) &= \left(x_L(n) + x_R(n)\right)/2 \\ y_C(n) &= \left(y_L(n) + y_R(n)\right)/2 \end{aligned} \tag{7}$$

when $n=\overline{1,N_S}$

7) Find the overlap between the segments. For each iteration great new matrix $dnc(y,x)$ Put overlapping segments in new matrix

$$\left(d_{nc}(y,x) \leftarrow b(y,x)\right) \tag{8}$$

And delete overlapping segments from $b(y,x)$ in the case of increasing the size of images can lead to a substantial increasing in the running time of this step and in the result is a Low speed algorithm as a whole.

8) Increase the counter for the loop by using the equation $n_C \leftarrow n_C + 1$

9) Check the conditions for the end of the loop. Exit from loop by doing any of the following conditions:

$$n_C = \hat{N}_C \tag{9}$$

$$N_S < \hat{N}_S$$

10) Combine the matrixes

$$\left(d_f(y,x) \leftarrow d_{nc}(y,x)\right) \tag{10}$$

according to this condition ($N_S > 6$) where $\hat{N}_S$ number of segments as explained above.

An example of the algorithm of selection texture regions on the image based on classification assessment density of contour elements. Fig. 3 shows the some of test of texture images and Contour filtering. images, distributed by several binary images; each one contains a homogeneous area points, lines, spots. For texture segmentation are used the method based on the selection texture regions on the image based on classification assessment density of contour elements. The goal of this method find a contour pixels of the input image, search the position of the contour elements and classify it's for different types (points, lines, and shapes) convert each region which had same contour to one segment, binary coding and mutual arrangement obtained polygon objects in the boundaries of the input image, segmentation resulting is formed as a code in the matrix.

As shown in Fig. 4, the result of tests image in Fig. 3(a, c), of The algorithm selection texture regions on the image based on classification assessment density of contour elements and the result of the method based on energy map.

In the table 1.The error of texture segmentation for each test image. The proposed method can reduce the average segmentation error to 14 times in comparison with the methods based on energy map.

Table 1. The values of errors in the texture segmentation of test images

Imge	Method	
	Proposed method	Method based on energy map
	An error in the texture segmentation for test images	
Cell; long line	0,0814	0
long line; spot	0,0501	0,0104
Cell ; point	0,0108	0
Point; short line	0,0344	0,0050

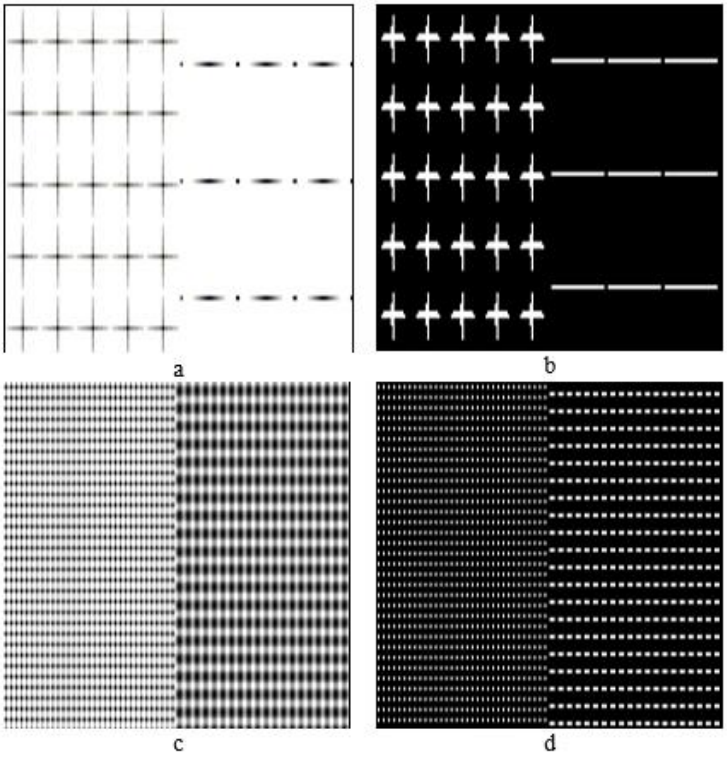


Fig. 3. Some tests images : a) cell, long line image ; b) contour for image (a); c) point, short line image; d) contour for image (c)

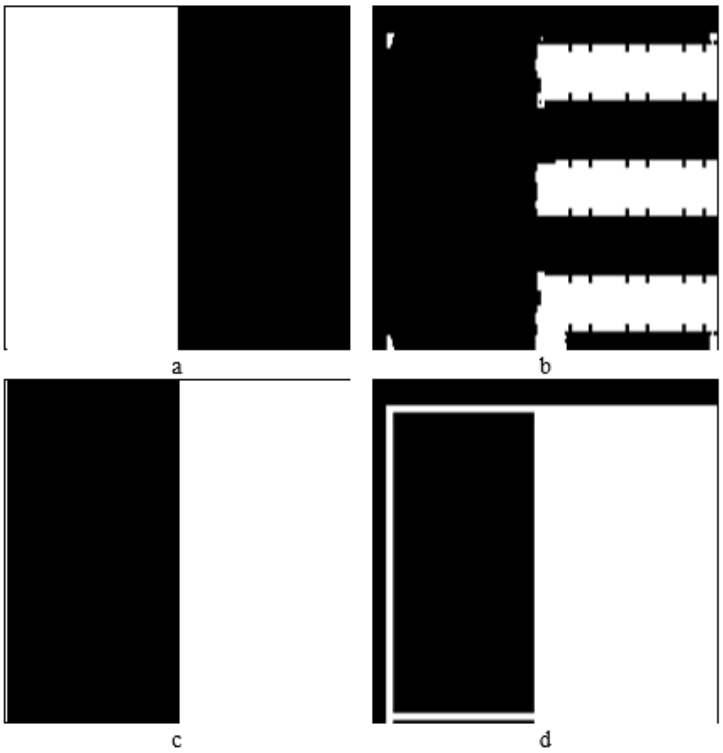


Fig. 4. a) segmentation result for image in Fig. 3(a) based on proposed method ; b) segmentation result for image in Fig. 3(a) based on energy map; c) segmentation result for image in Fig. 3(c) based on proposed method; d) segmentation result for image in Fig. 3(c) based on energy map;

In the table 1.The time of texture segmentation for each test image. The proposed method increase the time for segmentation to 18 times in comparison with the methods based on energy map.

Table 2. The values of errors in the texture segmentation of test images

Imge	Method	
	Proposed method	Method based on energy map
	Time of texture segmentation for test images	
Cell; long line	4,7793	1,7508
long line; zone	12,8403	1,7606
Cell ; point	15,1225	1,7458
Point; short line	129,473	1,7128

Summary. The algorithm selection texture regions on the image based on classification assessment density of contour elements, based on a search of the position of the contour pixel in the image and classify it's to different types. The advantages of propose method is:

4 Texture analysis or texture classification: define the contour element in the image and locate its as binary boundaries, classify theses contour elements for different type (point, line, cell, shape).

5 Texture segmentation: define the homogeneous regions, it helps to search different size of homogeneous regions and segment it as shown in the experimental above. The algorithm selection texture regions on the image based on classification assessment density of contour elements can reduce the average segmentation error to 14 times in comparison with the methods based on energy map and increase the increase the time for segmentation to 18 times in comparison with the methods based on energy map.

References

[1]. M. Kumar, K. K.Mehta," A Texture based Tumor detection and automatic Segmentation using Seeded Region Growing Method", IJCTA, Vol 2 (4), 855-859, July- August 2011.

[2]. Yong, J. Texture Analysis and Classification With Linear Regression Model Based on Wavelet Transform/ J. Yong, Z. Wang // IEEE TRANSACTIONS ON IMAGE PROCESSING. – 2008. – Vol. 17, № 8. – P. 1421–1430.

[3]. Akl, A. Structure Tensor Regularization for Texture Analysis / A. Akl, J. Iskandar // Image Processing Theory, Tools and Applications (IPTA). – Orleans. – 2015. – P. 592 – 596.

[4]. Boulkenafet, Z. Face Spoofing Detection Using Colour Texture Analysis/ Z. Boulkenafet, J. Komulainen, A. Hadid // IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY. – 2016. – Vol. 11, № 1. – P. 1 – 13.

[5]. Al zakki, H. M. Texture image segmentation based on classification of contour elements and logical addition of classes / H. M. Alzakki, V. Yu. Tsviatkou // 2016 Al-Sadeq International Conference on Multidisciplinary in IT and Communication Science and Applications (AIC-MITCSA, IEEE). – 2016. – P. 1 – 6.

EXPERIENCE IN ORGAANIZING EDUCATIONAL PROCESS IN BIG DATA ANALYTICS AT BSUIR

M. BATURA, Doctor of Engeneering Sciences
Rector BSUIR, Full Professor, Member of the International Higher Education Academy of Sciences, Honored Worker of Education of the Republic of Belarus

S.K. DZIK, PhD
First BSUIR vice-rector, PhD, associate professor

B. ZIBITSKER, PhD
President and CEO BEZNext, Emeritus professor of BSUIR

D. LIKHACHEVSKY, PhD
BSUIR Faculty of Computer-aided Design Dean, PhD, associate professor

I. TSYRELCHUK, PhD
BSUIR Lifelong and E-learning Studies Faculty Dean, Information and Computer-aided Systems Design Department Head, PhD, associate professor

K. YASHIN, PhD
BSUIR Engineering Psychology and Ergonomics Department Head, PhD, associate professor

The Belarusian State University of Informatics and Radioelectronics, Republic of Belarus
E-mail: sdick@bsuir.by, bzibitsker@beznext.com, likhachevskyd@bsuir.by, tsyrelchuk@gmail.com, yashin@bsuir.by

Abstract. For the last three years BSUIR has been offering Big Data educational courses. The initial course was conducted by Dr. Boris Zibitsker and Dr. Dominique Heger remotely. One part of the course included theory and the second one was Capstone project. Many students faced the problem of understanding the English materials and so the consequent courses were read in Russian. We always look for the way to improve educational process. Currently we use IBM's Big Data University Ambassodar program and students have a remote access to IBM's Big Data clusters to complete exercises. Fourth year students of Computer-aided Design Faculty took introductory courses and currently there are 58 3rd year students studying Fundamentals of BIG DATA, Introduction to data analysis and R technology, and Machine learning algorithms.

The classes are conducted in Russian and the students complete the exercises at the time suitable for them. This approach of conducting classes in Russian and completing exercises using IBM resources improved the process significantly. We are working on expending Big Data curriculum and offering courses to the students and PhD candidates of different faculties.

Introduction. The demand of modern production, the tasks of the banking sector, the development of nuclear energy and rapidly developing scientific research - all these and many other factors necessitate the training of specialists with higher education in the field of BIG DATA. In order to meet these requirements, as well as pass ahead of the development of the national economics and science, the Belarusian State University of Informatics and Radioelectronics started looking for solutions to the problems of training specialists in Big Data Analytics several years ago.

In this presentation we will share our experience in organizing the educational process in the Big Data Analytics and the plan of expanding the program at different departments and PhD programs.

Training of the specialists for working with large amounts of information - Big Data Analytics

Here are the main steps taken by the University for solving the problem of specialists with higher education training over the past 3 years.

1) There was organized scientific and technical cooperation with BEZNext (USA) specializing in the field of BIG DATA.

2) Specialized scientific and practical conference BIG DATA and Advanced Analytics was organized to be held annually. This conference enables specialists to discuss the development and application of modern information technologies in this area.

3) Commercial courses were set up. The courses provide teaching in English for those who want to learn the basic approaches and technologies of BIG DATA.

4) A supercomputer suitable for processing large amounts of information was purchased.

5) Master students of the Faculty of Computer Systems and Networks conduct research in the field of BIG DATA as a part of Master's theses.

6) Practice-oriented Master course of the Faculty of Computer Systems and Networks opened a new specialty for studying BIG DATA technologies.

7) Business relations with IBM employees (USA), who developed the educational program Ambassador and organized BIG DATA University to study modern technologies for processing large amounts of information, were established.

8) Three teachers were trained to conduct classes on BIG DATA in Russian with students of the Computer-aided Design Faculty.

9) Finally, the classes aimed at studying the main technologies of BIG DATA are organized for the 4th year students of Computer-aided Design Faculty.

The study of BIG DATA technologies was organized in the autumn semester of 2016 for the students of the 4th year (a group of 25 people). Students are trained in the field of Information Systems and Technology (in ensuring industrial safety), the qualification obtained is a system engineer, and the period of studies presupposes 4 years of full-time education.

Together with the covering of two planned IT disciplines 1) System software and 2) Modern programming languages, students were offered the Big Data Analytics materials designed with the practical experience of the IBM BDU program.

The Ambassador program and BDU are organized by IBM specialists. The founders of the

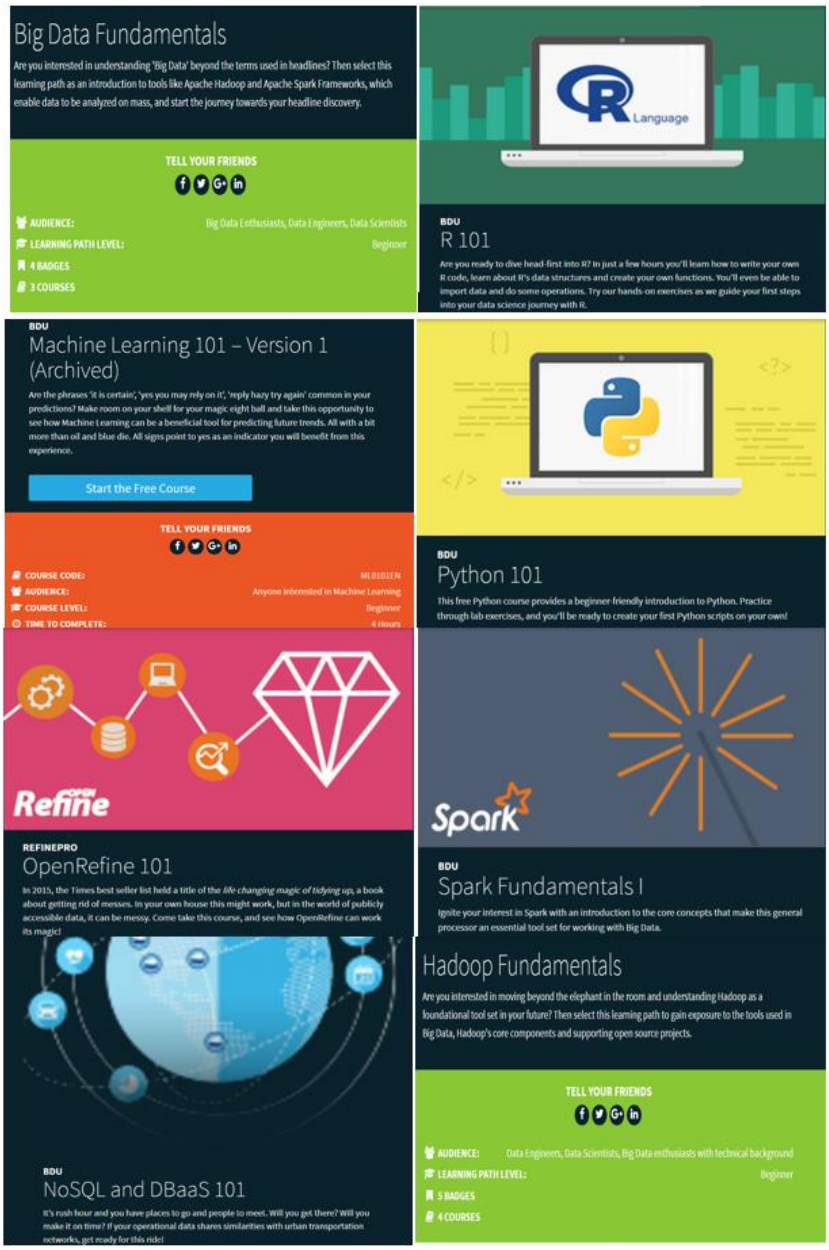
noted educational programs are IBM specialists working in the USA, Canada, India and other countries.

Universities of the USA, China, and Russia have been using educational resources of IBM Big Data University (BDU); BSUIR joined the BDU programs at the end of 2016.

In July-August 2016 three teachers of the Engineering Psychology and Ergonomics Department (BSUIR), which trains system engineers in two specialties: 1) Engineering and psychological support of information technology and 2) Information systems and technologies (in ensuring industrial safety), began the preparation of BIG DATA training materials for working with the students.

The educational program Ambassador BDU includes eight following sections (modules).

1) BIG DATA Fundamentals
2) Introduction to Data Analysis Using R
3) Machine Learning Algorithms
4) Python
5) Introduction to Open Refine
6) Spark Fundamentals
7) NoSQL
8) Hadoop Fundamentals



The following three sections (modules) were selected as the primary steps for mastering the

BDU materials by IBM in Russian with the students of the 4th year:

1 The Basics of BIG DATA
2 Introduction to the analysis of data using R technology
3 Algorithms of machine learning

As already mentioned above, in June 2016 three teachers were assigned to train students. All the three teachers completed their master's and postgraduate studies and are fluent in English which is necessary for covering BDU educational materials provided by IBM. The beginning of the classes with students was scheduled for November 1, 2016.

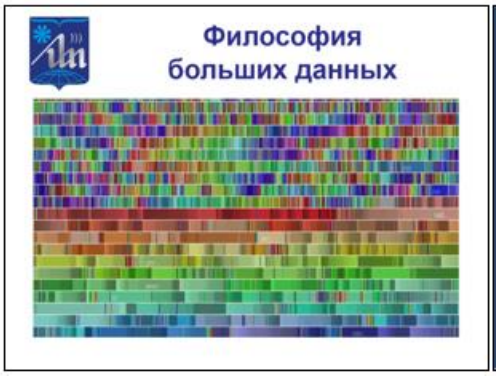
Teachers studied the BDU materials in July-August 2016 and in September-October they developed lectures and practical classes for students in Russian. In preparing the training materials the teachers used not only BDU resources (provided by IBM), but also studied educational resources on relevant topics at other universities around the world.

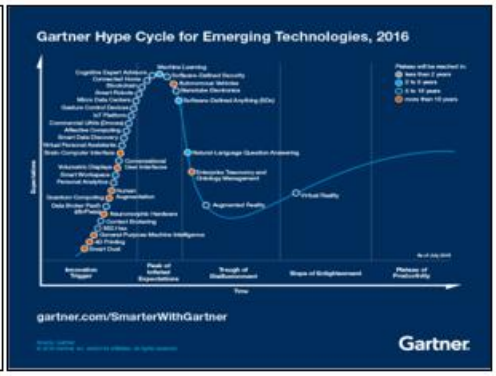
One of the teachers completed the training, passed exams and received certificates on the following 6 courses in July-August 2016:

1 BIG DATA Fundamentals (BDU от IBM)
2 Introduction to R-DataCamp Course (BDU от IBM)
3 Introduction to Machine Learning (DataCamp)
4 Intermediate R (DataCamp)
5 Intro to Statistics with R: Correlation and Linear Regression
6 Hadoop Fundamentals I (BDU от IBM)

Further on, we will consider the questions the teachers gave to the students in the process of studying the individual sections (modules) of BIG DATA.

Section "BIG DATA Basics". For the theoretical study of the section "Fundamentals of BIG DATA" by the students 18 hours of lectures (9 lectures) were allocated.





The section included the covering of the following 8 questions:

1. Definition and description of big data; its role in the economics and in the activity of enterprises.

2. Stages of big data technologies development.

3. Holistic approach to the big data development.

4. Examples of the big data formation: data and indicators of sensors in production; information from social media sources; results and arrays of scientific research results; information coming up with the expansion of data warehouses.

5**.** Explanation of the reasons for forming an integrated big data Platform for a general merger of what, otherwise, would be separate information stores with their own separate and autonomous analytics; in other words, the reason why the integrated Big Data Platform requires the connection of all types of information, individual before, into a single unit and a powerful information space (the creation of an information lake).

6. Identification of the importance of data management that will subsequently ensure big data control.

7. Description and determination of the components necessary for the formation of the Big Data Platform.

8. Comparison and confrontation of the following concepts: inactive data processing (data-at-rest processing); data stream processing (data-in-motion processing); data warehouses processing; contextual search.

Upon mastering the section "Fundamentals of BIG DATA" students passed the exam successfully. After the local BSUIR exam students were able to log in at BDU by IBM and cover the discipline "BIG DATA 101" and to get BDU certificates (examples of certificates are presented below).

Section "Introduction to data analysis using R technology". For students' studying this section 24 hours of lectures and 36 hours of practical training were allocated.

The content of the discipline is as follows:

Topic 1 Introduction to R Technology
- § 1.1 Numerical vectors
- §1.2 Factors
- §1.3 Arithmetic operations

Topic 2 Simple manipulations: numbers and vectors
- §2.1 Vector arithmetic
- §2.2 Logical vectors
- §2.2 Symbol vectors

Topic 3 Objects, their modes and attributes
- §3.1 Attributes, Modes
- §3.2 Getting and installing

Topic 4 Ordered and unordered factors
- §4.1 Obtaining factors from a categorical variable
- §4.2 Obtaining factors from a quantitative variable

Topic 5 Arrays and Matrices
- §5.1 Product
- §5.2 Transportation
- §5.3 Matrix Tools
- §5.4 Linking Arrays

Upon mastering the section "Introduction to data analysis using R technology" students successfully passed the exam. After the local BSUIR exam students were able to log in at BDU by IBM and cover the discipline "BIG DATA 101" and to get BDU certificates (examples of certificates are presented below).

Section "Algorithms of machine learning". For the students studying this section 24 hours of lectures and 36 hours of laboratory work were allocated.

The content of the discipline is as follows:

Topic 1 Introduction to data analysis and machine learning. Logical classification methods
- §1.1 Examples of machine learning application
- §1.2 Problems of retraining
- § 1.3 Python for data analysis
- §1.4 Working with vectors and matrices in NumPy
- §1.5 Decision trees

Topic 2 Metric and linear classification methods
- §2.1 Nearest neighbors algorithms
- §2.2 Parzen window method
- §2.3. Emissions detection
- §2.4 The stochastic gradient method
- §2.5 The SAG algorithm

Topic 3 Support vector method. Logistic regression
- §3.1 The essence of support vectors method
- §3.2 Application of support vectors method
- §3.3 The essence of logistic regression
- §3.4 Application of logistic regression

Topic 4 Linear regression. Dimension reduction and principal components method
- §4.1 Singular decomposition
- §4.2 Crestal regression
- §4.3 The LASSO Method
- §4.4 Approach to the characteristics selection

Topic 5 Compositions of algorithms. Neural networks
- §5.1 Bagging and Random Forest
- §5.2 Gradient boosting
- §5.3 Back propagation

Topic 6 Clustering and visualization
- §6.1 Lowering the dimension
- §6.2 Solving semi-supervised learning tasks

Topic 7 Machine Learning in Applied Problems
- §7.1 Working with numerical characteristics
- §7.2 Categorical and textual features
- §7.3 Data preprocessing

Prospects. In future, in spring semester (April and May 2017), two new groups of students will be trained in BIG DATA basics under the IBM Big Data University (BDU) program. The first group will consist of 32 3rd year students specializing in Engineering and psychological support of information technology (system engineers); the second group is represented by 26 students of the 3rd year specializing in Information systems and technologies (in industrial safety) (system engineers). Both

groups study 4 years full-time.

Together with the development of the planned IT discipline "Virtual Reality Technologies" the students of the first group will be offered the materials of the same three sections (modules) of BIG DATA developed last semester. These are: 1) The fundamentals of BIG DATA; 2) Introduction to data analysis using R technology; 3) Algorithms of machine learning. 24 hours of lectures and 32 hours of practical training are allocated to master all these three sections (modules). This is supposed to be a good experience for teaching in consolidating and improving the material, although with other students.

The students of the second group will be offered the same three sections (module) for parallel learning, but with two other IT disciplines already: 1) System software and 2) Modern programming languages. 24 hours of lectures and 18 hours of practical classes will be allocated for covering the section (module) "Fundamentals of BIG DATA". The two other modules will require the total 12 hours of lectures and 12 hours of laboratory work.

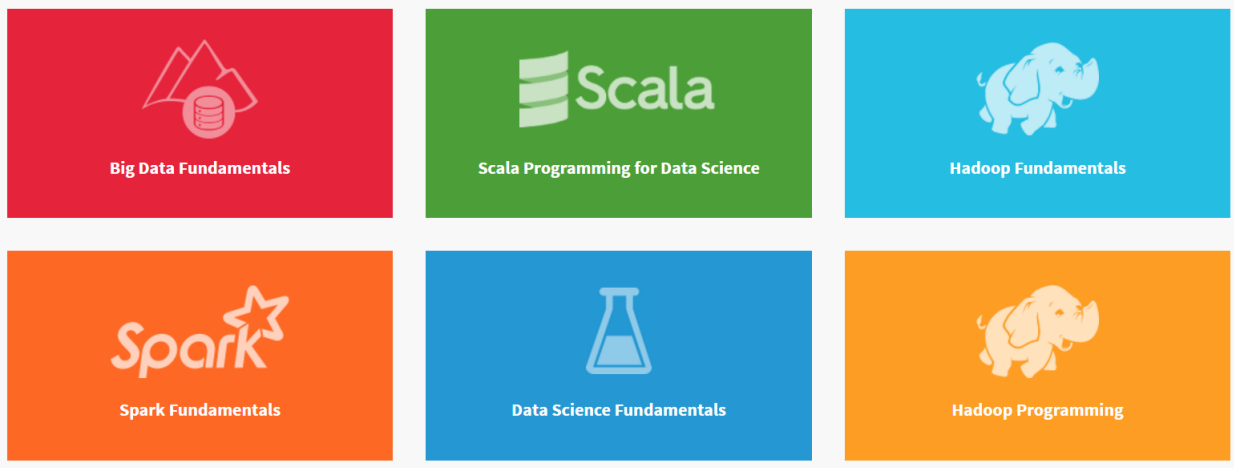
The number of sections (modules) is expected to be increased from 3 to 5 in the 2017-2018 academic year. The teachers will prepare two more new modules: 1) Python; 2) Introduction to Open Refine.

The following key technologies are required to prepare a high-level specialist for BIG DATA: Kafka, Spark, Storm, R, Cassandra, NewSQL, Columnar Database, In Memory, NoHadoop, NoSQL, OLTP, OLAP, ERP, Map Reduce, Scala, etc. Teachers pay special attention to the preparation of practical classes following the recommendations of the IBM BDU educational program and the educational programs for BIG DATA developed by other universities.

When we started Big Data program, it was a new field. Today, 3 years later, there is nearly no University teaching Computer Science, Business Administration, Finance or Economics with the students unaware of applying Big Data analytics to solving practical tasks.

Most of the requests for new projects financing by new startups include plans for Big Data analytics application. BSUIR invested a lot of efforts and energy into the introduction of Big Data Analytics education. Its success will depend on implementing this technology into the teaching courses at BSUIR and applying this technology by all PhD students in their theses.

We are planning to establish partnership with several vendors in open source projects using Machine Learning Algorithms. This will provide the students with the access to the state-of-the-art technology and prepare them to the challenging opportunities in the future.



NERVOUS CRISIS IN GAMERS UNDER THE INFLUENCE OF COMPUTER MEDIA

A. DAVIDOVSKI, PhD
Associate Professor, the Chair of Engineering Psychology and Ergonomics, BSUIR

K. KARANEUSKI
Research Scientist, BSUIR

K. MEZIANAYA
Research Scientist, BSUIR

K. YASHIN, PhD
Head of the Department of Human Engineering and Ergonomics, BSUIR

The Belarusian State University of Informatics and Radioelectronics, Republic of Belarus
E-mail: kira.m_2010@mail.ru

Abstract. It was established that 36.2% of respondents used to spend 40 or more hours a week in the virtual world. The majority of the respondents, 90.8%, is fond of computer games and devote to it about 55% of the time on average (100% is the total time spent at the computer). 78 students (54.9%) wake up due to fear and anxiety. The analysis showed that the number of people to wake up due to fear and anxiety among the users who combine computer games with watching movies is credibly higher: the criterion $\chi 2 = 5.83$. The article presents the dream description of one of the respondents.

Keywords: addiction (psychology), computers, gambling, dreams, mental health.

Computer online games, social networking, movies are an important part of the today's entertainment industry. Numerous studies have shown that both Internet in general and games in particular have a significant impact on the mental health of users (Yashin, Mezianaya, Zalivaka & Karaneuski, 2013; Seyyed, Mohammad, Fereshte & Mehdi, 2011).

Scientists have already established personal and behavioral deviations typical for people with Internet-addicted (Avetisova, 2011; Egorov, Kuznetsova & Petrova, 2005; Koronczai et al., 2011; Koc & Gulyagci, 2013). Other studies have found a relationship between behavioral disorders in computer-addicted persons and the symptoms of mental disorders (Dalbudak et al, 2013; Goldsmith, Keck, Khosla, McElroy, 2000). However, the problem of impact of the virtual world images on the human psyche, including unconscious processes has been hither to poorly study. In this regard, Freud's direction is especially up-to-date: "The study of dreams may be regarded as the most trustworthy approach to the exploration of the deeper psychic processes"(1922).

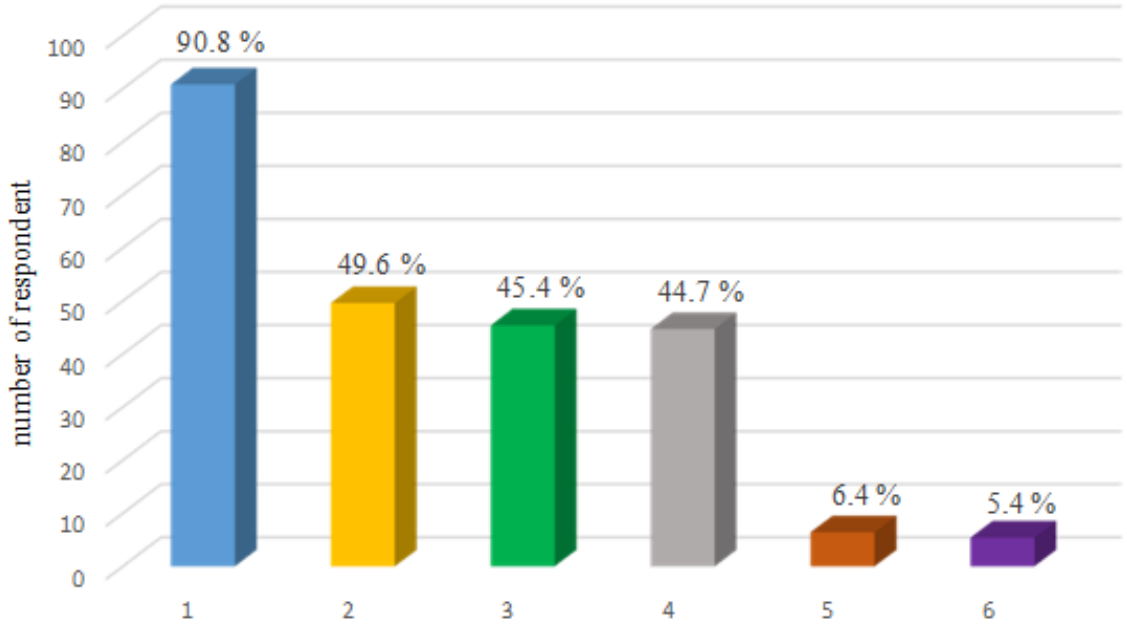
The objective of this study is to analyze the combined impact forms of the virtual world on the health of students.

Method of the Study. A survey using a continuous questioning method was conducted among 142 students studying information technology at one of the technical universities of Minsk, all of them 2-5-year students, 119 (84.4%) being males and 23 (16.2%) females. The average age of the respondents was 19.7 years. The average duration of the virtual world use was 9 years. The survey was conducted in the second half-year (in April, 2015). All respondents gave an informed consent to participate in the study. In order to study the influence of the virtual world on the health of the students a special questionnaire developed by K. Mezianaya, K. Yashin and K. Karaneuski entitled, "Method of Screening Diagnostics of Computer Addiction and its Effect on Physical and Mental Health" was used. The questionnaire included questions on the time spent in the virtual world, the structure of sleep and the nature of dreams. Six forms of virtual reality was analyzed: computer games, social networking, surfing, watching television series and movies, stock gambling and gambling (cards),

cybersex. The students were allowed to mark more than one answer. To analyze the time devoted to computer activity a period of 168 hours (7 days) with gradation in three intervals: from 1 to 24 hours, 25 to 39 hours, 40 hours or more was taken. Insomnia were identified by questions about waking up with fear, anxiety and unpleasant dreams, since according to the results of the studies conducted the relation between sleep disorders and anxiety disorders was established (Remizevich, 2010; Remizevich, 2013; Strygin, Yumatov, Levin, 2010).

In this study, an analysis was made of the frequency distribution of symptoms of pathological involvement in computer activities and sleep disorders. Contingency tables were made for the two samples, and $\chi 2$ criterion was calculated for the symptom of waking up due to anxiety and fear. The data obtained in this study were processed with standard application Microsoft Office Excel 2010 and a package of STATISTICA 10.0.

Results of the Study. Students were engaged in a full variety of forms offered by the virtual world. Two virtual world forms were used by 54 persons (38%), three forms by 48 persons (33.8%), and four by 15 persons (10.6%), while two students used all six forms. At the time of the study, 22 students (15.5%) used one form: 13 persons played computer games, 5 persons used social networks, and the rest spent their time surfing and watching movies. The analysis showed that the majority of users, 128 persons (90.8%) were fond of computer games.



1-computer games; 2-surfing; 3-social networks; 4-viewing television serials;
5-stock gambling; 6-cybersex

Fig. 1. Distribution of the virtual world forms per the frequency ofuse byrespondents, %

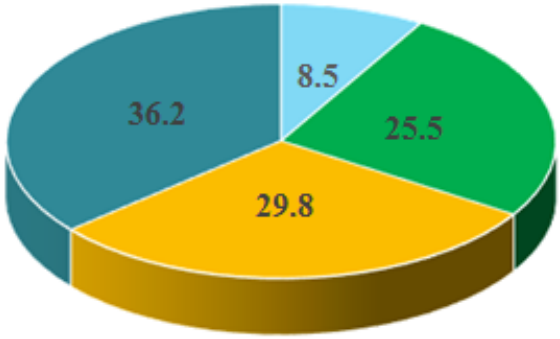
As shown in Fig. 1 computer games take the first place by the number of users in the group studied. Three virtual world forms are equally popular: watching movies, surfing (including video hosting) and social networks.

Analysis of the time spent in the virtual world per day showed that an average of 55% of the time is devoted to computer games (of 100% being the total time spent at the computer). 46.8% of respondents dedicate to it from 80 to 100% of the time.

Fig. 2 shows that 51 students (35.9%) stay in the virtual world more than 40 hours a week. Extra long stay, from 80 to 168 hours, took place in 12 students (8.5%). Their interest has acquired an addictive nature: they either do not visit the virtual world at all or stay there around the clock for seven days at a stretch, engaging mainly in computer games. Currently, scientists believe that one

reliable sign of computer addiction is when the duration of regular participation in virtual space as a sheer pastime without performing work exceeds 38 hours per wee (Young, 1998).

The majority of respondents play different genres of games, but they have preferences in the choice of action strategies of characters. The analysis was made with regard to psychological motives of computer game addiction.



from 1 to 24 hours – 25.5% of respondents; 25-39 hours – 36.2% of respondents;
40-79 hours – 29.8% of respondents; 80 hours and more – 8.5%

Fig. 2. Distribution of respondents by duration of stay in the virtual world during a week, %

As seen in Table 1, striving for suppression and subjugation of others in the games attracted 5.6% of respondents more than the hero action for rescue and protection. This fact testifies to the choice of aggressive plots by a large part of the gamers. As shown by the analysis, the gamers who want to achieve superiority and to rescue or protect others in the games, eagerly used violence and destruction in the virtual world to attain this aim. Devotion to orientation games may be explained by the fact that such psychological process as orientation in space and understanding of complex logical-and-grammatical structures and tasks are basically common due to localization in the parieto-occipital (low parietal) regions of the left hemisphere (Luria, 2006).

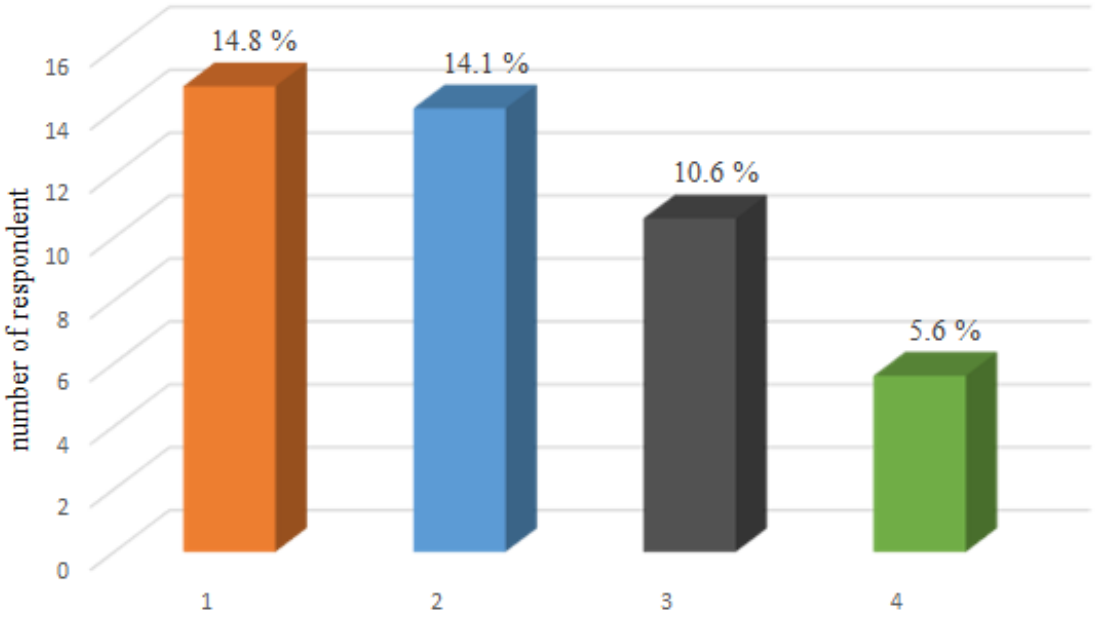
Table 1. Distribution of answers according to factors that characterize psychological motives of computer games addiction

Factor	Answer option	Number of answers	%
The aim of actions in the game of the virtual image controlled by the gamer	Rescue and protection	46	32.4
	Destruction and suppression	54	38.0
	Orientation in a new environment	47	33.1
	Achievement of superiority	68	47.9

Analysis of the impact of the virtual world on the unconscious processes in students. The destructive nature of the influence of the virtual world on the psyche of the students is supported by the fact that in 29 cases (20.4%) they stated to themselves in a dream: “I just dream it!”. 82 students (57.7%) used to wake up due to unpleasant nature of the dream. The analysis showed that 78 students (54.9%) woke up involuntarily due to fear and anxiety. In general, fear, anxiety, and unpleasant nature of dreams are the cause of waking up in 101 respondents (71.1%). At the same time, 45 persons (32%)

mainly rate their dreams as pleasant. This confirms the psychoanalytical fact that despite the negative experiences in dreams, the fulfillment of desires occurs in them. Freud pointed out that in the course of the dream some unconscious desires may be implemented: "… the painful elements of our daily thoughts are able to force their way into our dreams only if at the same time they are able to disguise a wish-fulfillment." He also argued that the theory of the anxiety-dream belongs to the psychology of neuroses: "… anxiety in dreams is an anxiety-problem and not a dream-problem"(1900/2012). On the other hand, nightmares (incubi), according to psychoanalysis data, represent a punishment for realizing something illicit, forbidden, or a breach of moral standards.

The study in the combined effect of several virtual world forms on the mental processes of users revealed a special role of computer games, along with the video films. The analysis showed that among users of these two forms there was a credibly higher number of people waking up due to fear and anxiety, criterion $\chi 2 = 5.83$, as compared with those not combining viewing video films with playing games.



Note: 1 - the feeling of one's selectness and special mission; 2 - appearance of strange dreams; 3 - presence of a hero of computer games; 4 - continuation of the game in a dream

Fig. 3. Analysis of the nature of dreams, %

The presence of game scenes in dreams occurs in 16.3% of those surveyed. As shown in Fig. 3, 8 students continue to play in dreams (5.6%), in 15 respondents (10.6%) a hero or a virtual world image were present in the dreams, owing to his/her identification with the characters of games. 21 respondents (14.8%) experienced strange premonitions in a dream: "strange sensation of future changes, presentiments of their selectness, some mission." Analysis showed that such students often played strategy genres more than 6 hours per day, and some of them for up to 168 hours a week.

Freud stated that dreams are a response to all that is relevant at this time to the sleeping soul and that "every dream without exception treats of oneself. Dreams are absolutely egoistic..." (1900/2012).

Dreams of a 18-year student:

A huge ship made of white metal is hanging in the sky. My friend is in the air, holding on to the edge of the board, so as not to fall. I'm trying to pull him out, but he falls down and naval guns destroy him in flight. I suddenly find myself among the crowd, which the naval guns start to shoot at.

And I just have to run faster than shells." He further said: "Today, my nightmares have fairly the same structure:

– There is a character in dreams to cause a sense of discomfort: fight with him, cross-talk, bullying and so on.
– Other characters of a dream are indifferent to his actions in relation to me;
– My attempts to resist this villain cause a general resentment of other characters in the dream;
– The very attempt to confront him turns out to have devastating consequences for me."

As established by psychoanalysis, such scenes are dreams of punishment of oneself. At the time of survey the student stated that his experience of video games was 14 years. He visited by about five virtual world forms, limiting the sessions to 4 hours. He believed that "everything was permitted" to the hero of the computer game and "he was feeling the impact of the virtual image on himself." In his dreams there was a sexual theme, dream in a dream, self-punishment and the sense of connectedness. The extent of his immersion in the game and/or the Internet is such that sometimes it leads to the loss of perception of the surrounding.

Conclusion. The results of the study indicate that some of the students studying technical specialist, are involved in all the variety of services offered by the virtual world. Stay in cyberspace for 40 or more hours per week is established in 36.2% of respondents, which can indirectly indicate a pathological passion.

Striving to achieve the desired results leads to the ignoring of morality standards in the games, resulting in the inner psychological conflict. The combined effect of video films and game plots destabilizes the mind of users, leading to disturbed sleep due to anxiety and fear in 54.9% of respondents. It is a sign of formation of anxiety and depressive disorders.

The changes in mental activity at an unconscious level in combination with psychopathological phenomena, developing under the influence of computer games, may pose a threat of uncontrolled behavior of such persons.

References

[1]. Seyyed S. A., Mohammad R. M., Fereshte J., Mehdi E. (2011). The effect of psychiatric symptoms on the internet addiction disorder in Isfahan's University students. Journal Research Medical Science, V. 16(6): 793-800.

[2]. Avetisova A.A. (2011). Psichologicheskie osobennosti igrokov v kompyuternye igry [Psychological features of computer gamers]. Psichologicheskie osobennosti igrokov v kompyuternye igry [Psychology Journal of the Higher Economic School]. V. 8, №4, 35-58.

[3]. Egorov A.U., Kuznetsova N.A., Petrova E.A. (2005). Osobennosti lichnosti podrostkov s Internet-zavisimost'yu [Personality characteristics of a teenager with the Internet dependence]. Voprosy psikhicheskogo zdorov'ya detey i podrostkov [Problems of mental health of childrens and teenagers], V. 5. № 2, 20-27.

[4]. Koronczai B., Urbán R., Kökönyei, Paksi B., et al. (2011). Confirmation of the Three-Factor Model of Problematic Internet Use on Off-Line Adolescent and Adult Samples. Cyberpsychology Behavior Social Networking. V.14. №11: 657–664.

[5]. Koc M, Gulyagci S. (2013). Facebook addiction among Turkish college students: the role of psychological health, demographic, and usage characteristics. Cyberpsychology Behavior Social Networking. V.16 (4):279-84.

[6]. Dalbudak E, Evren C, Aldemir S, Coskun KS et al. (2013). Relationship of internet addiction severity with depression, anxiety and alexithymia, temperament and character in university students. Cyberpsychology Behavior Social Networking. v.16 (4):272-278

[7]. A. Goldsmith TD, Keck P.E, Jr, Khosla UM, McElroy S.L. (2000). Psychiatric features of individuals with problematic internet use. Journal of Affective Disorders. v. 57(1-3):267–72.

[8]. Freud S. (1922). "Beyond the Pleasure Principle." Retrieved from http://www.bartleby.com/276/1.html.

[9]. Remizevich R.S. (2011). O reciprocnyh vzaimootnosheniyakh trevozhnykh rasstroistv i naruchenyi sna [On reciprocal interrelations of anxiety disorders and sleep disturbances]. / Remizevich R.S., G.P.Kostyuk//

Actual'nye voprosy somnologii [Topical somnology problems]. – Moscow p.58.

[10].Remizevich R.S. (2013). Insomnicheskie narusheniya pri trevozhnyh rasstroistvakh u voennosluzhashchich molodogo vozratas extremalnyh vidov professional'noy deyatel'nosti [Insomnia disturbances in cases of anxiety disorders in young servicemen involved in extreme professional activities]. Avtoreferat na soiskanie uchenoy stepeni candidata medicinskikh nauk [Abstract for the degree of Candidate of Medical Sciences]. St. Petersburg.

[11]. K.M.Strygin, E.A.Yumatov, Ya.I.Levin. (2011). Sootnoshenie lichnostnyh osobennostey i characteristic nochnogo sna cheloveka [Interrelation of personal features and characteristics of human nocturnal sleep] // Aktual'nye voprosy somnologii [Topical somnology problems]. – Moscow. – 63.

[12]. Young K.S (1998): How to Recognize the Signs of Internet Addiction and a Winning Strategy for Recovery. New York, Publisher Wiley . -256 p.

IMPACT OF CULTURE AND TRANSFER OF EMBRYOS IN MICE: ASSESSMENT OF MICROARRAY

N. KARAGENÇ, MD, PhD ***K. ESMEN[2]*** ***G. DOĞAN[2]*** ***L. KARAGENC[2]***

Assistant Professor,
Pamukkale University,
Medical Faculty, Department of Medical Biology

Pamukkale University, Faculty of Medicine. Dept. of Medical Biology, Turkey
[2]Adnan Menderes University, Veterinery Faculty, Dept. of Histology, Turkey
E-mail: nkaragenc@hotmail.com

Keywords: Mouse, innate immunity, fetus, lung, embryo, in vitro embryo culture, oxygen.

The aim of the present project was to test whether or not in vitro culture of mouse zygotes under atmospheric oxygen concentrations has any different effects on the expression different genes. Two control groups and one experimental group were included in the project. Control group consisted of fetuses obtained through mating of female mice that were not super-ovulated. The experimental group consisted of fetuses generated by transfer of in vitro-developed blastocysts obtained through in vitro culture of zygotes. In both groups, fetuses obtained on day 18 of gestation were weighed and expression genes in fetal lung tissue samples were determined using micro-array analysis. Results indicated that fetal weight was significantly reduced in experimental groups. Micro-array and qRT-PCR analyses demonstrated that when compared to fetuses in the Control-1 group, the level of 13 mRNA transcripts were significantly reduced in Experimental group. Taken together, data gathered from the present study indicates that in vitro embryo culture and embryo transfer leads to reduced fetal weight, delayed lung development. Evidence gathered from these studies is important as it leads to a better understanding of cellular/molecular mechanisms underlying increased susceptibility of newborns with low birth weight to various allergic/infectious diseases.

SOFTWARE FOR PREDICTING THE RELIABILITY OF THE ELECTRONIC SYSTEM BY ITS TECHNICAL STATES SET ANALYSIS METHOD

N.I. TSYRELCHUK
Master of Information and Computer Systems Design BSUIR

S.S. DZIK
Master of Information and Computer Systems Design BSUIR

S. BOROVIKOV, PhD
Associate Professor, Department of Information and Computer Systems Design BSUIR

I. TSYRELCHUK, PhD
BSUIR Lifelong and E-learning Studies Faculty Dean, Information and Computer-AIDED Systems Design Department Head, PhD, associate professor

V. KAZIUCHITS
5-th year student of the Faculty of Computer-Aided Design BSUIR

N. ZHIDILIAEVA
Foreign Languages Department №1 teacher, BSUIR

E. SHNEIDEROV M.Eng.
Department of Information and Computer Systems Design BSUIR

The Belarusian State University of Informatics and Radioelectronics, Republic of Belarus
E-mail: tsyrelchuk@bsuir.by, sdick@bsuir.by, bsm@bsuir.by, vladisgenerator@gmail.com, nzhidilyaeva@gmail.com

Abstract. The method of direct selection of the system technical states is the simplest and most understandable method for predicting the reliability of an electronic system under any connection scheme from the point of view of the reliability of its component parts (devices). If the number of devices entering the system is more than 8 ... 10, algorithms are needed to generate and further process large amounts of data on possible technical conditions of the system. Such algorithms are proposed and used in the developed software tool for assessing (predicting) the reliability of the electronic system. This software tool allows you to build a structural scheme of reliability in an interactive mode on a computer. After entering the data about the system components (devices) reliability the computer calculates the reliability indicator automatically.

In a number of cases technical systems, including electronic systems of different functional purpose, from the point of view of reliability, have such a structure of their component parts connection (or interaction) that is not reduced to parallel-sequential or sequentially parallel circuits. An example of such a connection structure is the bridge circuit (Figure 1).

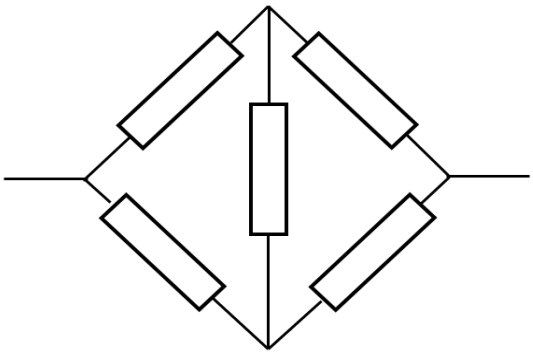


Fig. 1. Bridge connection of electronic system devices in terms of reliability

In practice, such schemes may exist for electronic systems containing information and computer subsystems.

We will assume that the system under consideration contains n devices. The system can be in two states: functionality and failure. We mark the state of the system by R symbol. We assume that R takes value 1 if the system is functional and value 0 if it fails. The state of the j^{th} device of the system is indicated by x_j symbol. We assume that x_j takes value 1 if the j^{th} device is working without fail, and the value is 0 if it fails (j = 1, 2, ..., n).

The state of the electronic system depends on the state of its devices, i.e.,

$$R = \varphi(x_1, x_2, \ldots, x_n), \tag{1}$$

where φ is the symbol of the functional connection.

Function (1) is called the structural function of the system. For existing electronic systems the following relations hold true:

$$\varphi(0, 0, \ldots, 0) = 0;$$

$$\varphi(1, 1, \ldots, 1) = 1;$$

$$\varphi(x_1, x_2, \ldots, x_n) \geq \varphi(y_1, y_2, \ldots, y_n) \text{ provided that x } x_j \geq y_j, (j = 1, 2, \ldots, n).$$

Physically, the last condition denotes that the failure of the device can not transfer the system from its inoperative into operable state.

In [1-3] one can become familiar with the methods of calculating and evaluating the reliability of technical systems not reducing to parallel-sequential or sequentially parallel schemes.

The simplest and most understandable method for calculating the probability of an operable state of these systems is the method of direct selection of system technical states. This method can be successfully applied also to reliability calculation of the systems that, from the point of view of reliability, reduce to parallel-series or series-parallel devices connection.

The essence of the direct search method. Taking into account the criterion for the failure of the electronic system the entire set of its technical states G is divided into two subsets: the operable states of G_1 and the inoperative states of G_0. For each state of the electronic system $X = \{x_1, x_2, ..., x_n\}$ we can calculate its probability p_x and then find the probability of a functional state of the electronic system ($P_{эс}$):

$$P_{\text{ЭС}} = P\{X \in G_1\} = \sum_{X \in G_1} p_X \quad , \tag{2}$$

where $P\{...\}$ hereinafter denotes the probability of the event indicated in curly brackets.

The probability of an inoperative state of the electronic system can be defined as

$$Q_{\text{ЭС}} = P\{X \in G_0\} = \sum_{X \in G_0} p_X \quad . \tag{3}$$

For the probability of the system X state, under the assumption of the independence of the devices from the point of view of the occurrence of their failures, the formula is true

$$p_X = \prod_{j=1}^{n} p_{x_j}, \tag{4}$$

Where p_{xj} is the probability of the state x_j of the j^{th} device of the system ($x_j = 1$ or $x_j = 0$).

In the general case, without applying IT-technologies, the method is justified if a number of devices in the system is relatively small ($n \leq 6 \ldots 10$), since with the number of devices in the system $n = 10$ the number of possible technical states S for the system is $S = 2^n = 1024$, which is actually problematic for engineering analysis.

To quantify the probability of the electronic system's functionality it is necessary to consider the possible technical states of the system. These states are determined by the technical states of the devices making up the system [4]. For devices, as a rule, one of two states can exist: either inoperative or operable, while for the system as a whole there are many states that differ by combinations of operability and inoperability of system devices. Some of these states correspond to the state of inoperability of the system as a whole, others - to the state of operability.

Estimating the reliability of a complex electronic system by examining the system as a whole in practice causes many difficulties due to the excessive number of the possible system S technical states. For example, with the number of devices $n = 30$, the value $S > 1$ billion. The total amount of data needed to describe such a number of possible technical system states will approach the size of Big Data [5]. However, with the value $n < 30 \ldots 40$ the reliability analysis of the electronic system can be performed using traditional methods on a computer with medium performance, but this requires algorithms employing prediction principles [6]. These principles allow us to systematize and further process large volumes of data on possible technical states of electronic system. Such algorithms are proposed and used in the developed application software for assessing (predicting) the reliability of the electronic system by the method of enumeration of technical states.

The developed software allows to build a structural scheme for calculating the reliability of the electronic system in the interactive mode on a computer. After entering the data about the reliability of the components (devices) the computer calculates the reliability index of the system automatically. The application software was developed at the Information and Computer Systems Design Department of BSUIR.

Below are the explanations that allow you to get the most general idea about the software. A fragment of its main window is shown in Figure 2.

The software application for the implementation of the project for the reliability of the electronic system analysis includes the following steps:

1 Clarification of the electronic system condition. At this stage it is necessary to find out what is the operability of the system: complete or partial (the latter presupposes normal functioning). Such operability is considered to be the performance of the electronic system while its functional parameters are within the limits specified in the technical documentation.

2 Structural scheme for calculating the electronic system reliability construction. This scheme is built by the user, with the help of computer graphic capabilities software and the electrical structural and / or functional circuit of the system under consideration accounting the conditions for its opera-

bility and normal functioning. The dialog box of structural scheme for calculating the electronic system reliability construction is shown in Figure 3. It shows the already constructed structural scheme for calculating the reliability of the electronic system under consideration. The software allows you to set the functional parts of the system interactively on the working field of the monitor, assign names (identifiers) to them and make the necessary connections taking into account the conditions of the electronic system operation.

3 Entering data about the components (devices) of the electronic system reliability and calculating the probabilities of its operable and inoperative states. The calculation is performed automatically by the software by means of analyzing the built structural scheme of the system reliability. The windows for data input, output of the results analysis and calculation of the system reliability indicators are shown in Figure 4.

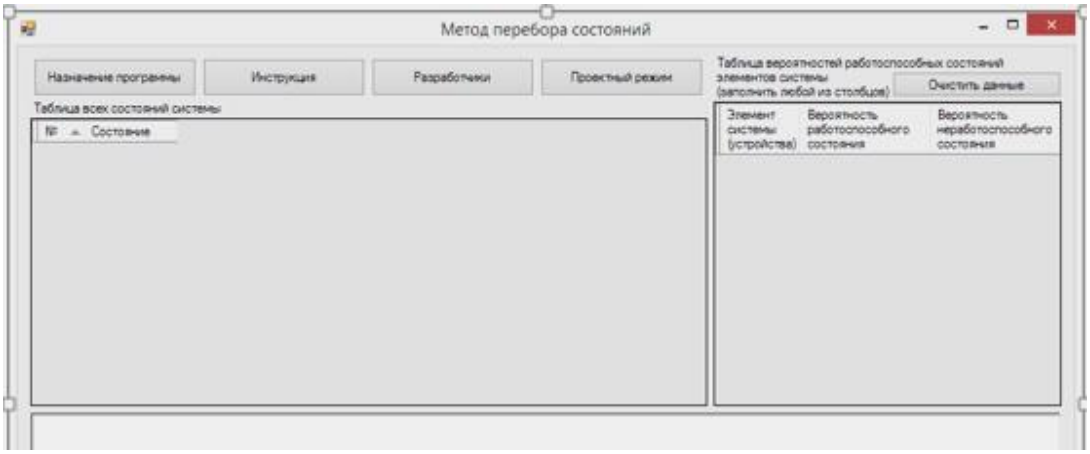


Fig. 2. Main window of the software

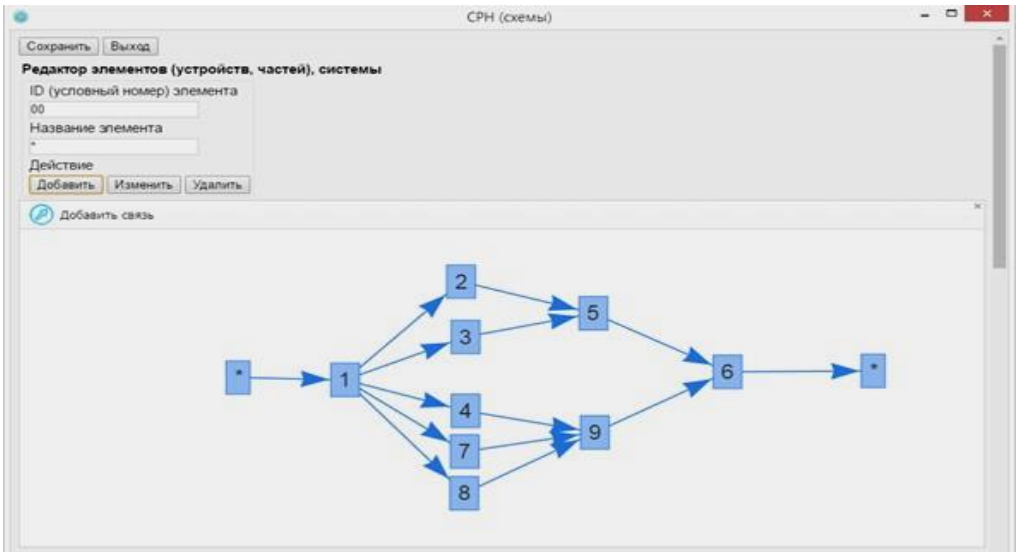


Fig. 3. The constructed structural scheme for calculating the system reliability

The right table of Figure 4 displays the entered data on the reliability of the system components (devices). In the left table the first column shows the system state number, the second column indicates which subset of technical states the state of the electronic system is related to: green color - a subset of operable states, red - a subset of inoperative states. In the following columns number "1" in the symbolic designation of the device status corresponds to its operable state, and digit "0" - to the inoperative state of the device. The rightmost column of the table indicates the probability of the

corresponding state of the electronic system with four digits after the decimal point. Value "0" (zero) means that the probability of this state is less than 0.00005. When the value is slightly more than 0.00005, rounding to the value 0.0001 is performed.

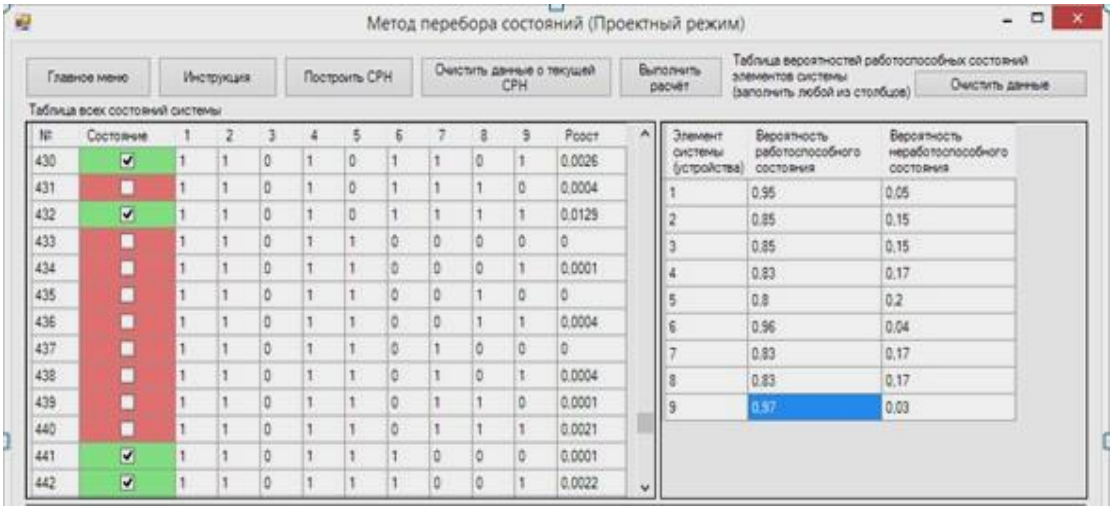


Fig. 4. Electronic system reliability analysis and calculation results

The bottom window (see Figure 4) shows the results of calculating the electronic system probabilities of an operational and inoperative state, indicating the number of states of each subset.

Developed software testing proved that it successfully solves the problem of estimating (predicting) the reliability of the electronic system with up to 30 functional parts (devices) in it. With their number n = 30 it took about 1.5 hours to complete the task. The computer had the following resources: RAM - 8 GB; Processor - Intel, 2 cores - 2.5 GHz.

Conclusion: the development of successful algorithms and their use for computer information processing allows to solve the problems in calculating the complex electronic systems reliability, even though the description of technical states of the hypothetical data volume is close to the size of Big Data.

References

[1]. Половко, А. М. Основы теории надёжности / А. М. Половко, С. В. Гуров. – 2-е изд., перераб. и доп. – СПб. : БХВ-Петербург, 2006. – 704 с.

[2]. Надёжность технических систем : справочник / Ю. К. Беляев [и др.] ; под ред. И. А. Ушакова. – М. : Радио и связь, 1985. – 608 с.

[3]. Шишмарёв, В. Ю. Надёжность технических систем : учебник для студ. высш. учеб. заведений / В. Ю. Шишмарёв. – М. : Изд. Центр «Академия», 2010. – 304 с.

[4]. Цырельчук, Н.И. Оценка надёжности электронной системы методом анализа множества её технических состояний / Н.И. Цырельчук, С. М. Боровиков, С. С. Дик, И. Н. Цырельчук // Современные средства связи: материалы XXI Междунар. науч.-техн. конф., 20–21 окт. 2016 года, Минск, Респ. Беларусь; редкол.: А.О. Зеневич [и др.]. – Минск: УО ВГКС, 2015. – С. 126–127.

[5]. Batura, M. Big Data Volumes and Some Approaches to the Creation of Corporate Analytical Systems / M. Batura, S. Dzik, I. Tsyrelchuk, S. Borovikov // BIG DATA and Advanced Analytics. Использование BIG DATA для оптимизации бизнеса и информационных технологий: сб. материалов II Междунар. науч.-практ. конф. (Минск, Республика Беларусь, 15–17 июня 2016 года) / редкол. : М.П. Батура [и др.]. – Минск : БГУИР, 2016. – С. 74–80.

Borovikov, S. Prediction in Big Data Technology / S. Borovikov, E. Shneiderov, N. I. Tsyrelchuk, S.S Dzik // BIG DATA and Advanced Analytics. Использование BIG DATA для оптимизации бизнеса и информационных технологий : сб. материалов II Междунар. науч.-практ. конф. (Минск, Республика Беларусь, 15–17 июня 2016 года) / редкол. : М.П. Батура [и др.]. – Минск : БГУИР, 2016. – С. 98–101.

SOFTWARE FOR EVALUATING THE ELECTRONIC SAFETY SYSTEM RELIABILITY IN CASE OF LARGE VOLUME OF DATA ABOUT ITS TECHNICAL CONDITIONS AVAILABILITY

S.S. DZIK
Master of Information and Computer Systems BSUIR

N.I. TSYRELCHUK
Master of Information and Computer Systems Design BSUIR

S. BOROVIKOV, PhD
Associate Professor, Department of Information and Computer Systems Design BSUIR

I. TSYRELCHUK, PhD
BSUIR Lifelong and E-learning Studies Faculty Dean, Information and Computer-aided Systems Design Department Head, PhD, associate professor

S.K. DZIK, PhD
First BSUIR vice-rector, PhD, associate professor

N. ZHIDILIAEVA
Foreign Languages Department №1 teacher, BSUIR

The Belarusian State University of Informatics and Radioelectronics, Republic of Belarus
E-mail: sdick@bsuir.by, tsyrelchuk@bsuir.by, bsm@bsuir.by, nzhidilyaeva@gmail.com

Abstract. Evaluation of the design reliability of a complex electronic system causes many difficulties. This is a consequence of an excessive number of possible technical states of the system. In a number of cases, the number of these states and the data volumes for their description is so large that they fall under the notion of Big data. As the result, the usual methods of processing such volumes of data are not possible. As a way out of the situation, we propose to use the simplification of analysis which is based on the decomposition of the system.

We have developed software for applying the decomposition method to assess the reliability of electronic security systems. It allows you to build a protected object (the building plan with its premises) in an interactive mode with the help of a computer, place the components of an electronic security system in the premises, allocate subsystems and perform their analysis in terms of the reliability and protection of the premises.

The reliability of a technical system is one of its most important properties. This property largely determines the success of the task assigned to the system. Therefore, when designing a technical system of any functional purpose, the question of predicting the index of its reliability is urgent [1]. Such indicator should be considered as the efficiency preservation coefficient C_{ep} or its modifications [2]. Coefficient C_{ep} in accordance with State Standard 27.002-89 is a generalized name of a group of indicators used in various industries and having their own names, designations and definitions [3]. For electronic security systems, it is appropriate to consider the probability of ensuring the

security of an object or an individual as such an indicator. The value of this indicator depends on both the reliability of the technical devices included in the system, and the probabilities of perception and/or the correct threat signals processing. The values of these probabilities are determined by the temporary failures of the system's devices, which are the consequence of the external environment influence (climatic factors, electromagnetic influences, etc.) on the system and its constituent parts [4, 5].

To quantify the probability with which the security of an object is ensured, it is necessary to consider the possible technical states of the system and to take into account the efficiency coefficients corresponding to these states. It is logical to use the probabilities of object protection as the efficiency coefficients (provided that the system is in this technical condition). Technical conditions of the system are determined by the technical states of the devices included in it [6]. For devices, as a rule, one of two states can exist: either inoperative or operable, while for the system as a whole there are many states that differ by combinations of operability and inoperability of system devices. Some of these states correspond to the state of inoperability of the system as a whole, others - to the state of operability. Depending on the combination of technical states (operable or inoperative), the functioning state of the electronic security system is characterized by different probabilities of object protection, or, it is said, different functioning efficiency.

Estimation of functioning efficiency of a complex electronic security system by considering the system as a whole in practice causes many difficulties due to the excessive number of possible technical states of the system S, which is defined as

$$S = 2^n, \tag{1}$$

where n is the total number of technical devices included in the electronic security system.

For example, in the case when a building contains 30 rooms and the installation of only one sensor on each entrance door and on each window (with one window in the room) is available, the number of possible technical states of the electronic security system will be

$$S = 2^{30+30} \approx 1{,}153 \cdot 10^{18}.$$

Which is important, this number takes into account only the sensors but not other devices of the electronic security system.

Considering that dozens of memory bytes are needed to store data on one technical state of the system, the total amount of data necessary to describe all possible technical states of the electronic security system can amount to a number that falls under the concept of Big Data [7].

Thus, the effectiveness of the electronic security system analysis is associated with the examination of a large amount of data about the system state. In this case it is impossible to process such a volume of data by traditional methods. The question arises, what is the way out of this situation, how to take into account the large amount of data on possible technical conditions of the electronic security system?

To solve the problem for engineering practice various methods for simplifying the analysis of system reliability can be proposed. One of these methods is decomposition [5, 8]. Its essence consists in dividing the system under consideration into smaller subsystems, each of which is much easier to analyze than the original system. Upon receiving reliability indicators of the subsystems it is relatively easy to find the reliability index of the system as a whole.

For analyzing the reliability of the electronic security system by the decomposition method application software was developed at the Information and Computer Systems Design Department of BSUIR. Below are the explanations that allow you to get the most general idea of this software.

The developed software allows to create a plan of the building in an interactive mode, to place sensors, video cameras and other security devices on the building premises, to allocate subsystems in the initial system on the basis of the composition and interaction of the technical devices of the system

consideration, i.e., in fact, perform the decomposition of the system, analyze the effectiveness of the of premises protection using the allocated subsystems, determine the efficiency (reliability) of the system as a whole. The probability of the premises protection (with the help of the electronic security system) from an offender's penetration is considered as an indicator of the effectiveness of the functioning of the system.

Figure 1 depicts the main window of the developed software. The menu bar at the top of the window is used to select the user's actions during preparation (configuration) and project execution. You can choose to execute a new project and save the results of its development and calculation, or open a previously saved project and continue its execution.

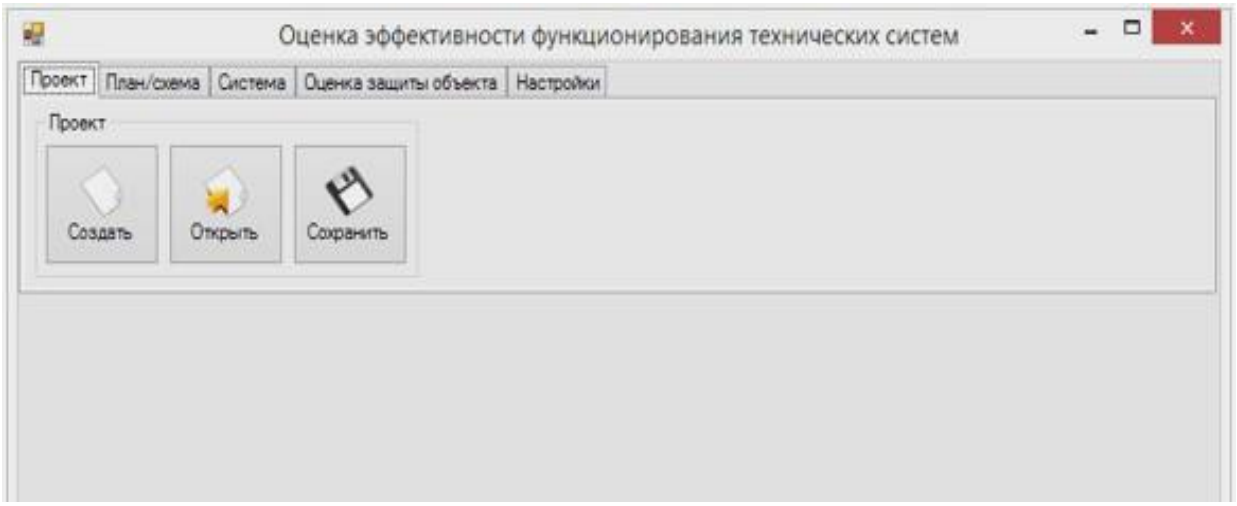


Fig. 1. Main window of the software tool

Figure 2 demonstrates a fragment of a constructed building plan, the premises of which will be protected by an electronic security system.

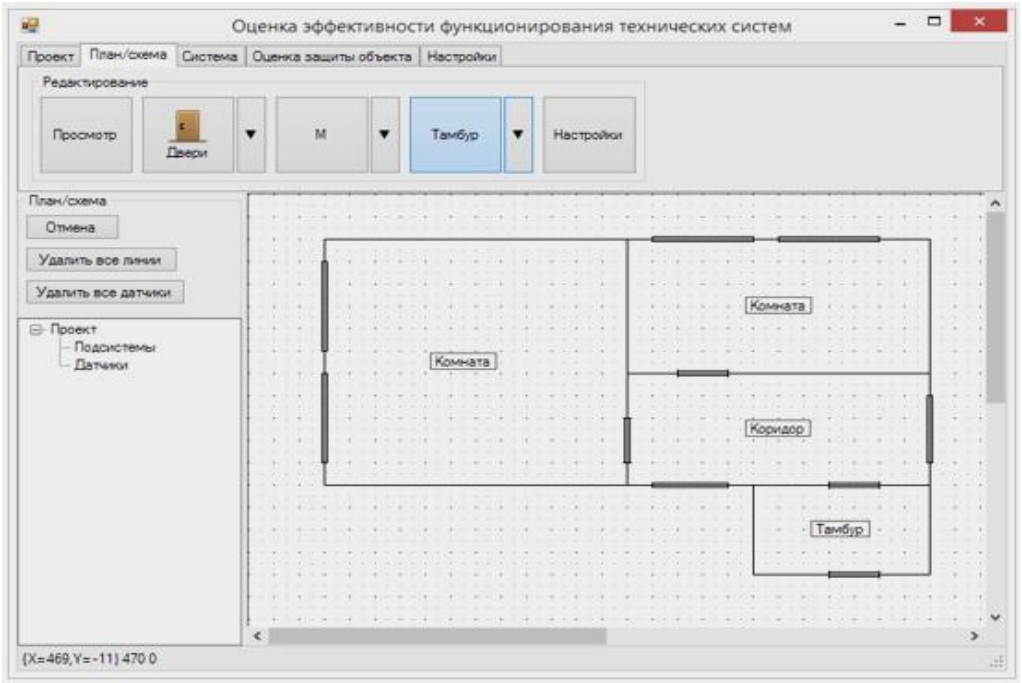


Fig. 2. Fragment of the protected building premises plan

The plan of the object (building premises, see Figure 2) is designed by a user with the help of the graphical capabilities of the software and the tools provided for editing the plan.

Figure 3 illustrates one of the options for the user to place sensors and other devices of the electronic security system on the built-up plan of the building's premises.

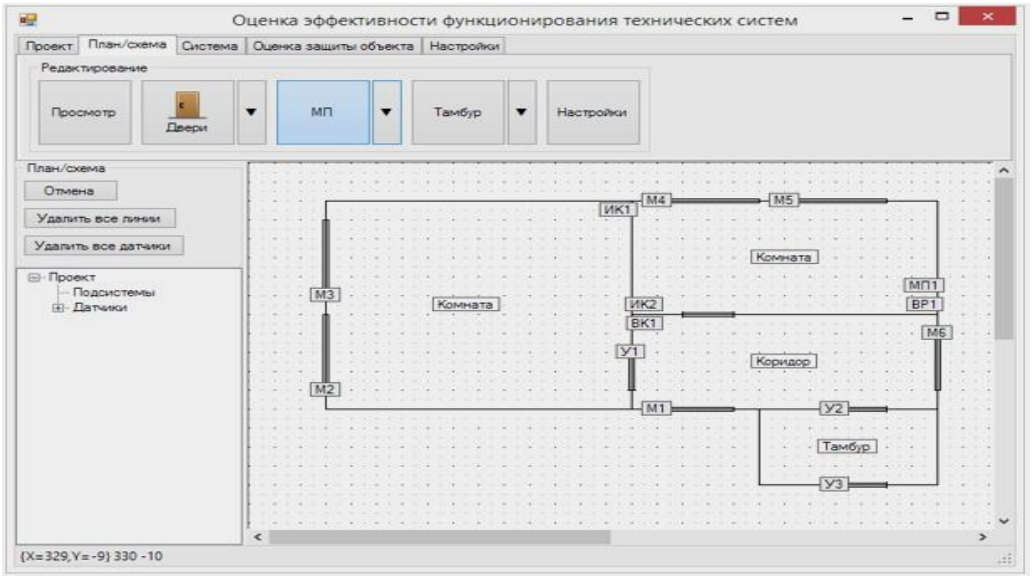


Fig. 3. Example of placement of system devices in the premises of a building

In Figure 3 the following designations are used: М1 ... М6 - magnetic contact sensors; У1 ... У3 - shock sensors; ИК1, ИК2 - infrared motion sensors; ВК1 - video camera; ВР - DVR; МП1 - microprocessor receiving and monitoring device.

Figure 4 illustrates the allocation of the subsystem, i.e., the actual implementation of the electronic security system decomposition.

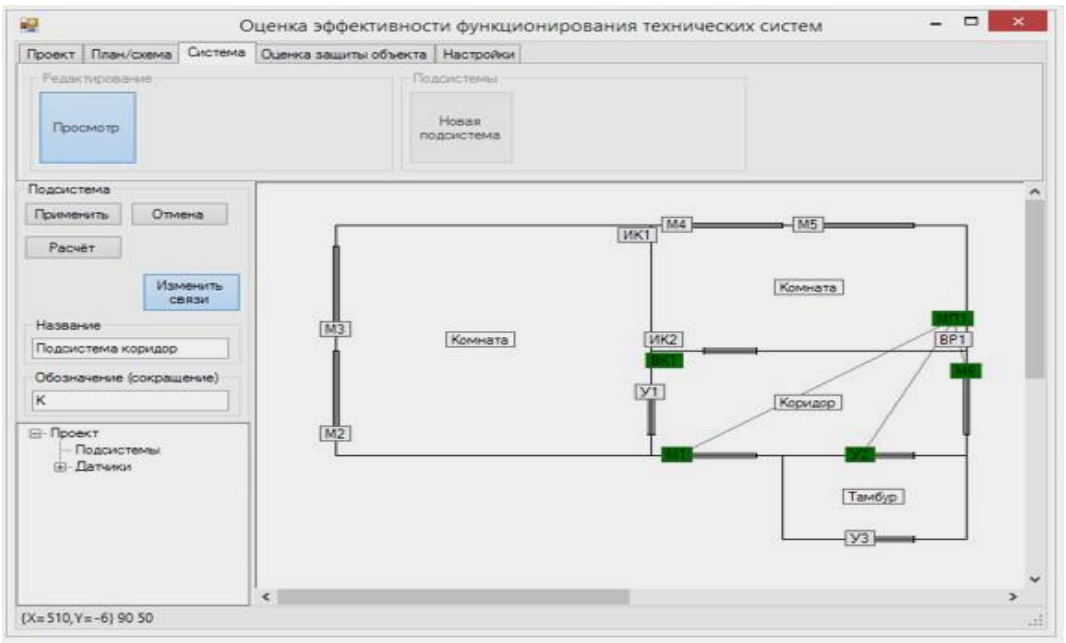


Fig. 4. Selection of the subsystem from the system devices

Let us describe the basic principles of decomposition with reference to electronic security systems. On the one hand, the subsystems should be distinguished from the condition of counteracting the infiltration of the intruder into the building's premises; on the other hand, they should be simple. The number of devices in the subsystems should not exceed 4 ... 6 units. The same subsystem device can be a part of two or more subsystems, for example, the МП1 device will be a part of all subsystems that will include at least one sensor, since the threat signals generated by the sensors arrive for processing at the device МП1 - microprocessor- control device. When real projects are implemented, the number of allocated subsystems having the same composition of devices and their interaction may turn out to be a remarkable number: dozens, sometimes hundreds.

The electronic security system functioning efficiency indicator (in the form of the probability of the object protection Pprot) is relatively simple to be obtained from the results of the subsystems performance analysis.

Thus, the use of complex technical systems decomposition makes it possible to isolate homogeneous information (to do system decomposition) from a large amount of data (Big Data) about possible technical states of a complex system and keep completing the task by traditional methods.

References

[1]. Borovikov, S. Prediction in Big Data Technology / S. Borovikov, E. Shneiderov, N. I. Tsyrelchuk, S.S Dzik // BIG DATA and Advanced Analytics. Использование BIG DATA для оптимизации бизнеса и информационных технологий : сб. материалов II Междунар. науч.-практ. конф. (Минск, Республика Беларусь, 15–17 июня 2016 года) / редкол. : М.П. Батура [и др.]. – Минск : БГУИР, 2016. – С. 98–101.

[2]. Надёжность в технике. Основные понятия, термины и определения. ГОСТ 27.002–89. – М. : Изд-во стандартов, 1990.

[3]. Состав и общие правила задания требований по надёжности. ГОСТ 27.003-90. – М.: Изд–во стандартов, 1991.

[4]. Боровиков, С. М. Теоретические основы конструирования, технологии и надёжности : учебник для инж.-техн. спец. вузов / С. М. Боровиков. – Минск : Дизайн ПРО, 1998. – 336 с.

[5]. Надёжность технических систем : справочник / Ю. К. Беляев [и др.] ; под ред. И. А. Ушакова. – М. : Радио и связь, 1985. – 608 с.

[6]. Цырельчук, Н.И. Оценка надёжности электронной системы методом анализа множества её технических состояний / Н.И. Цырельчук, С. М. Боровиков, С. С. Дик, И. Н. Цырельчук // Современные средства связи: материалы XXI Междунар. науч.-техн. конф., 20–21 окт. 2016 года, Минск, Респ. Беларусь; редкол.: А.О. Зеневич [и др.]. – Минск: УО ВГКС, 2015. – С. 126–127.

[7]. Batura, M. Big Data Volumes and Some Approaches to the Creation of Corporate Analytical Systems / M. Batura, S. Dzik, I. Tsyrelchuk, S. Borovikov // BIG DATA and Advanced Analytics. Использование BIG DATA для оптимизации бизнеса и информационных технологий: сб. материалов II Междунар. науч.-практ. конф. (Минск, Республика Беларусь, 15–17 июня 2016 года) / редкол. : М.П. Батура [и др.]. – Минск : БГУИР, 2016. – С. 74–80.

[8]. Дик, С.С. Прогнозирование эффективности функционирования электронной системы при наличии большого объёма данных о её технических состояниях / С.С. Дик, С. М. Боровиков, Н. И. Цырельчук, С.К. Дик // Современные средства связи: материалы XXI Междунар. науч.-техн. конф., 20–21 окт. 2016 года, Минск, Респ. Беларусь; редкол.: А.О. Зеневич [и др.]. – Минск: УО ВГКС, 2015. – С. 377–378.

ПОЛУЧЕНИЕ И АНАЛИЗ БОЛЬШИХ ОБЪЕМОВ ВИБРОМЕТРИЧЕСКИХ ДАННЫХ И СИГНАЛОВ

П.Ю. Бранцевич
Доцент кафедры программного обеспечения информационных технологий БГУИР, кандидат технических наук, доцент

Е.Н. Базылев
Ассистент кафедры программного обеспечения информационных технологий БГУИР

С.Ф. Костюк
Заведующий лабораторией систем вибродиагностики БГУИР

Белорусский государственный университет информатики и радиоэлектроники, Республика Беларусь
E-mail: branc@bsuir.edu.by

Abstract. Vibration sensors installed on a set of similar devices produce large volume of data for the analysis over long period of time. Accurate analysis of data allows to identify technical malfunctions at an early stage and efficient measures for their elimination. This paper shows examples of data analysis of vibration monitoring systems that control operation of turbine units

Введение. Для обеспечения достаточно комфортных условий существования человека требуется надежная, безаварийная и, при этом, экономически эффективная работа многих сложных, материалоемких, энергопотребляющих, большеразмерных технических объектов и производств. На таких объектах эксплуатируется большое количество дорогостоящего оборудования: двигатели, турбины, генераторы, насосы, компрессоры, вентиляторы. В ходе их работы осуществляется контроль разнообразных параметров, по которым можно судить об их исправности и работоспособности. При этом, параметры вибрации являются одними из важнейших и подлежат обязательному, в том числе и непрерывному, контролю для многих механизмов и агрегатов роторного типа, в основу механического функционирования которых положено вращательное движение [1-6].

Системы непрерывного контроля (мониторинга) определяют параметры вибрации в точках установки датчиков через небольшие промежутки времени (от нескольких секунд до долей секунды) и реагируют на возникновение аномальных ситуаций, проявляющихся в изменении вибрационного состояния, путем выработки сигналов управления устройствами сигнализации и защитного отключения. В качестве основных параметров вибрации для принятия таких решений используется среднее квадратическое значение (СКЗ) виброскорости и амплитуды ряда частотных составляющих вибросигнала, кратных частоте вращения вала (ротора) [1,7,8].

Однако факт возникновения ситуации, требующей останова технического объекта, во многих случаях имеет неоднозначное отображение в параметры вибрации. Стандартизованные критерии защиты отражают наиболее общие взаимосвязи [6,9], полученные на основе длительного опыта эксплуатации и исследования механизмов с вращательным движением, и далеко не всегда в полной мере могут удовлетворить эксплуатирующий и управляющий персонал.

Системы вибрационного контроля и защиты, построенные на базе компьютерной техники, позволяют реализовать разнообразные и сложные алгоритмы защиты, ориентированные на конкретные типы дефектов и ситуаций. Это, в свою очередь, позволяет избежать необоснованных («ложная тревога») срабатываний защитного отключения и не допустить «пропуска

дефекта».

Реализован и прошел апробацию на ряде турбоагрегатов алгоритм защитного отключения по вибрации, в котором учитывается несколько факторов:

− низкочастотная составляющая вибрации;

− оборотные составляющие вибрации;

− высокочастотная составляющая вибрации.

Сигнал на защитное отключение контролируемого механизма вырабатывается в том случае, если он выработан по одному из указанных критериев, или по нескольким критериям одновременно [10].

Также в результате функционирования компьютерных комплексов вибрационного контроля и мониторинга накапливаются большие объемы данных, содержащих информацию об изменении во времени различных вибрационных параметров для всех точек контроля [11,12]. Получаемые таким образом данные в определенной степени соответствуют современной концепции «больших данных». Они позволяют выявить изменение технического состояния контролируемого объекта или провести анализ причин, приведших к неисправности или отказу оборудования. Гораздо более полную информацию об эксплуатируемых механизмах содержат непрерывные вибрационные сигналы, регистрируемые на протяжении длительных временных интервалах и в разных режимах работы. Они в полной мере являются «большими данными» [13].

Обработка данных вибрационного мониторинга. Лабораторией систем вибродиагностики БГУИР разработан и внедрен на основных генерирующих электростанциях Беларуси компьютерный измерительно-вычислительный комплекс (ИВК) серии «Лукомль», реализующий функции вибрационного контроля, защиты и мониторинга турбоагрегатов [8,11,12,14]. ИВК решает задачи текущего штатного вибрационного контроля и защиты [6,9], а также создает суточные файлы, в которых сохраняется большой объем данных, представляющих изменение параметров вибрации в единицах виброскорости во времени (частота вращения вала, СКЗ виброскорости в частотной полосе 10-1000 Гц, значения амплитудных и фазовых параметров до десяти спектральных составляющих вибрации, кратных частоте вращения (порядковый анализ), пик-фактор исходного сигнала).

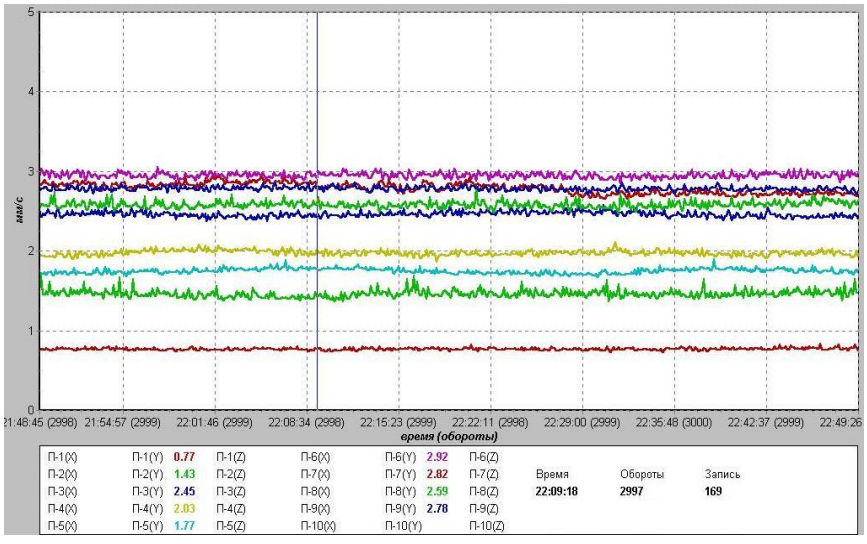


Рис. 1. Тренд СКЗ виброскорости (вертикаль) подшипниковых опор турбоагрегата при его нормальной работе

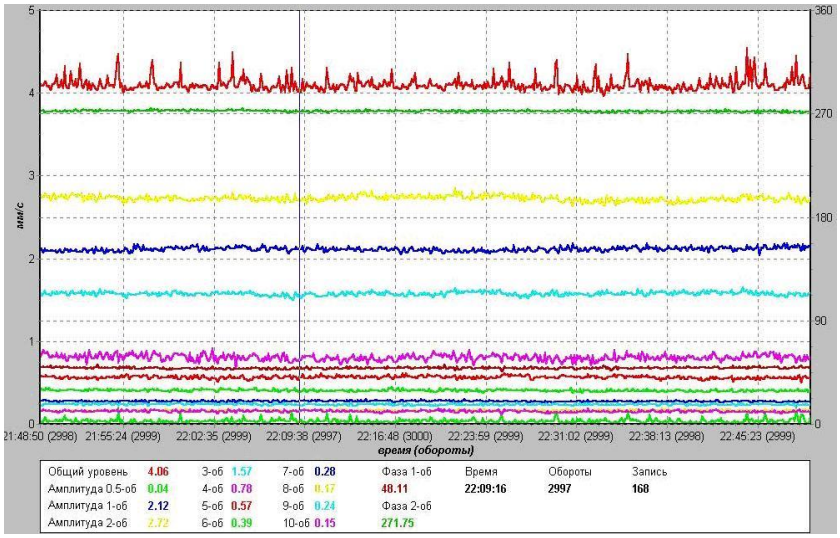


Рис. 2. Тренд СКЗ виброскорости (горизонталь), оборотных составляющих, фазы первой и второй оборотных составляющих для подшипниковой опоры возбудителя турбоагрегата

На рисунках 1-2 представлены примеры изменения во времени (примерно час), параметров вибрации, зафиксированных при контроле вибрации подшипниковых опор турбоагрегата. Даже визуальный анализ трендов параметров вибрации позволяет обнаружить их возможные изменения и, соответственно, изменение технического состояния контролируемого объекта.

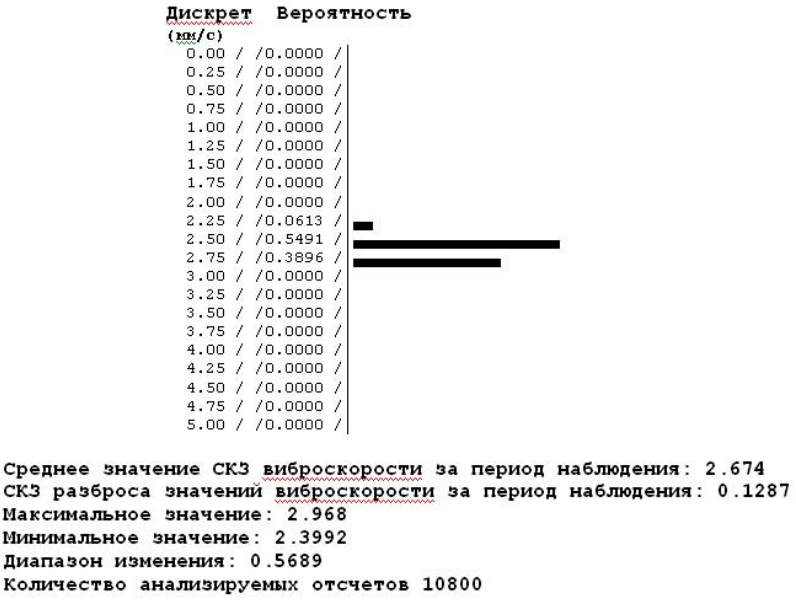


Рис. 3. Результаты обработки суточного временного тренда СКЗ виброскорости подшипника генератора при его нормальной работе

Однако его проведение связано со значительными временными затратами технического специалиста, к тому же, конечно хотелось бы получить какие-то численные оценки обрабатываемых данных. Поэтому актуальна автоматизация этой процедуры.

Статистическая обработка данных, получаемых ИВК «Лукомль», может быть выполнена с помощью достаточно простого программного средства. На рисунке 3 показан пример такой обработки, в результате которой получена гистограмма распределения исследуемого пара-

метра по уровню и численные значения, определяющие отличительные особенности его изменения. Эти вычисленные значения можно принять в качестве вектора информативно-значимых параметров для системы поддержки принятия решений по оценке изменения технического состояния контролируемого объекта. В качестве примера для сравнения на рисунке 4 приведены изменения СКЗ виброскорости подшипниковых опор турбоагрегата при возникновении дефекта, а на рисунке 5 результаты статистической обработки одного из этих параметров. Амплитудный диапазон для построения гистограммы выбирается с учетом реального вибрационного состояния контролируемого объекта и нормативных требований по уровню вибрации.

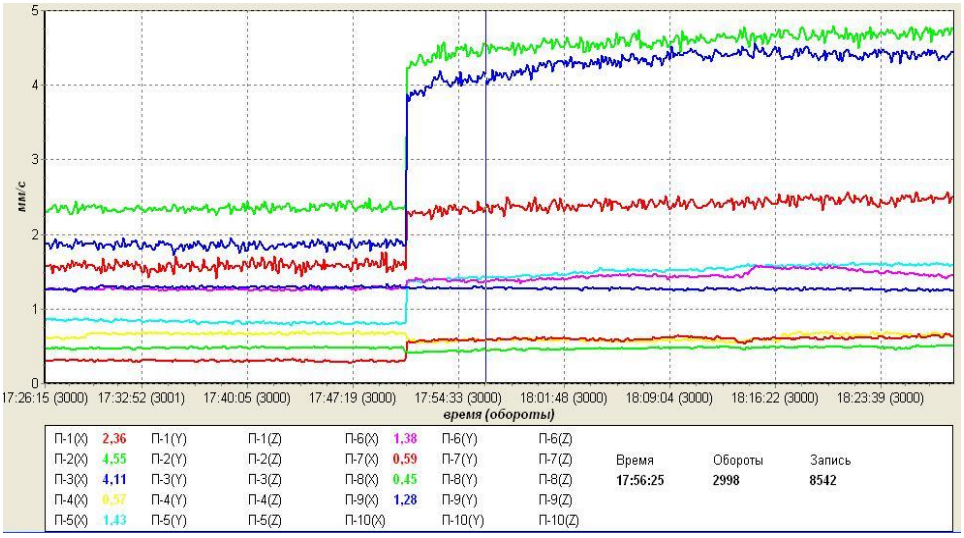


Рис. 4. Тренд СКЗ виброскорости (горизонталь) подшипниковых опор турбоагрегата при возникновении дефекта

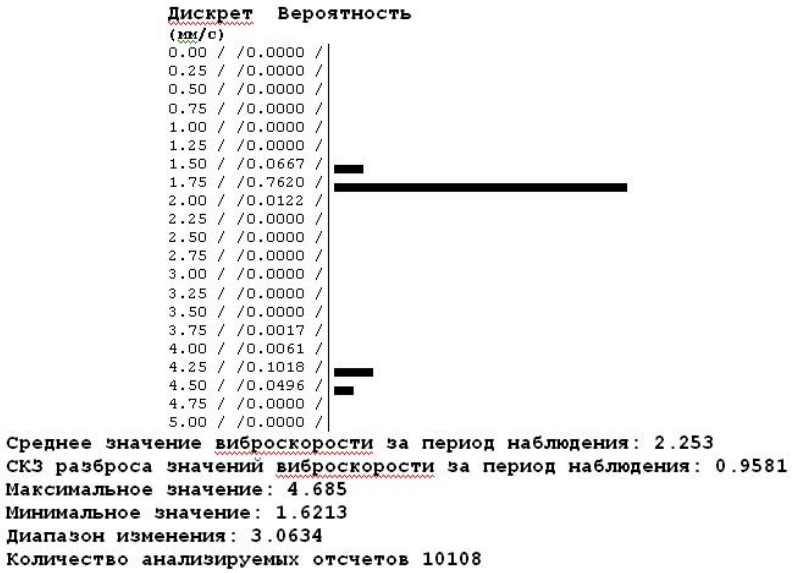


Рис. 5. Результаты обработки суточного временного тренда СКЗ виброскорости подшипника турбоагрегата при возникновении дефекта

Анализируя полученные результаты можно сделать ряд выводов.
При нормальном функционировании механизма:

– параметр «СКЗ разброса значений виброскорости за период наблюдения» имеет значительно меньшее значение по сравнению со «Средним значением СКЗ виброскорости за период наблюдения»;

– «Среднее значение СКЗ виброскорости за период наблюдения» попадает в квантиль максимальной вероятности гистограммы распределения;

– «Диапазон изменения» меньше, чем «Среднее значением СКЗ виброскорости за период наблюдения» (хотя при наличии случайных выбросов, помех или кратковременных значительных изменениях режима это условие может не выполняться);

– форма гистограммы распределения по уровню для анализируемого параметра представляет собой несколько, расположенных по соседству друг с другом, значащих квантилей.

При возможном дефекте или неисправности механизма:

– параметр «СКЗ разброса значений виброскорости за период наблюдения» сравним по величине со «Средним значением СКЗ виброскорости за период наблюдения»;

– «Среднее значение СКЗ виброскорости за период наблюдения» во многих случаях не попадает в квантиль максимальной вероятности гистограммы распределения;

– «Диапазон изменения» сравним или даже превышает «Среднее значением СКЗ виброскорости за период наблюдения»;

– форма гистограммы распределения по уровню для анализируемого параметра может иметь произвольный вид, причем значащие квантили могут располагаться с разрывом по амплитудной шкале.

Таким образом, предварительно проведенная статистическая обработка временных трендов параметров вибрации может значительно облегчить работу технического специалиста.

Непрерывные вибрационные сигналы несут еще больший объем информации. Длительные наблюдения за контролируемыми объектами позволили обнаружить кратковременные вибрационные всплески-возмущения [15]. Исследование таких возмущений представляет собой достаточно большую проблему, так как они носят случайный характер, а временные интервалы между ними могут составлять часы и даже сутки.

Необходимость обнаружения редких кратковременных изменений структуры вибрационных сигналов и последующее выявление причинно-следственных связей между их появлением и развитием дефектов, требует создания систем, способных накапливать и обрабатывать непрерывные вибрационные сигналы, отражающие вибрационное состояние объекта, на протяжении длительных временных интервалов. Такие системы могут быть построены по принципу распределенного сбора и централизованной обработки [16]. При этом подходе функции непрерывного ввода и регистрации на смарт-карте вибросигналов выполняют автономные, энергонезависимые устройства [17], а их обработка осуществляется на производительных вычислительных машинах, в том числе с неоднородными и распределенными ресурсами.

*Заключение***.** В результате сбора данных от вибродатчиков, установленных на множестве однотипных устройств, в течение длительного периода получается большой объем данных для анализа. Тщательно проведенная обработка этих данных создает предпосылки для выявления технических проблем на ранней стадии и их оперативного разрешения.

Литература

[1]. Неразрушающий контроль. Справочник. Том 7. Книга 2. Вибродиагностика /Ф.Я. Балицкий и др. М.: Машиностроение, 2005. – 485 с.

[2]. Ширман, А.Р. Практическая вибродиагностика и мониторинг состояния механического оборудования / А.Р. Ширман, А.Б. Соловьев. – Москва, 1996. – 276 с.

[3]. Барков, А.В. Мониторинг и диагностика роторных машин по вибрации / А.В. Барков, Н.А. Баркова, А.Ю. Азовцев. – СПб. : Изд. центр СПбГМТУ, 2000. – 169 с.

[4]. Bently, D.E. Fundamentals of Rotating Machinery Diagnostics/ D.E. Bently, C.N. Hatch, B. Grissom. – Canada.: Bently pressurized bearing company, 2002. – 726 pp.

[5]. Гольдин, А.С. Вибрация роторных машин / А.С. Гольдин. М.: Машиностроение, 1999. –344 с.

[6]. ГОСТ ИСО 10816–1–97. Вибрация. Контроль состояния машин по результатам измерений вибрации на невращающихся частях. Часть 1. Общие требования. – Введ. 1999–07–01. – Минск. Межгосударственный совет по стандартизации, метрологии и сертификации: ИПК Изд-во стандартов, 1998. Стандартинформ, 2007. – 18 с.

[7]. Бранцевич, П.Ю. Организация и опыт применения систем вибрационного мониторинга и защиты / П.Ю. Бранцевич, С.Ф. Костюк // Достижения физики неразрушающего контрол: сб. научн. тр. / Под ред. Н.П. Мигуна – Мн.: Институт прикладной физики НАН Беларуси, 2013. – с. 67-74.

[8]. Бранцевич, П.Ю. ИВК «Лукомль -2001» для вибрационного контроля /П.Ю. Бранцевич // Энергетика и ТЭК. −2008. − № 12 (69), −с. 19−21.

[9]. ГОСТ 25364–97. Агрегаты паротурбинные стационарные. Нормы вибрации опор валопроводов и общие требования к проведению измерений. – Введ. 1999–07–01. – Минск. Межгосударственный совет по стандартизации, метрологии и сертификации: ИПК Изд-во стандартов, 1998. Стандартинформ, 2011.– 12 с

[10]. Brancevich, P. Organization of the vibration-based monitoring and diagnostics system for complex mechanical system / P. Brancevich, X. Miao, Y. Li // 20th International Congress on Sound and Vibration. Bangkok, Thailand, 7-11 July 2013. − Curran Associates, Inc., NY 12571 USA, –pp. 612-619.

[11]. Бранцевич, П.Ю. Решение задач вибрационного контроля, мониторинга, оценки технического состояния механизмов и турбоагрегатов с помощью компьютерных комплесов / П.Ю. Бранцевич, С.Ф. Костюк, Е.Н. Базылев // Доклады БГУИР. − 2015. − № 2 (88). − с. 148-152.

[12]. Бранцевич, П.Ю. Компьютерный вибрационный мониторинг механизмов и турбоагрегатов/ П.Ю. Бранцевич, С.Ф. Костюк, Е.Н. Базылев // Доклады БГУИР. − 2015. − № 7 (93). − с. 5-10.

[13]. Фрэнкс, Б. Укрощение больших данных: как извлекать знания из массивов информации с помощью глубокой аналитики / Б. Фрэнкс; пер. с англ. А. Баранова. − М.: Манн, Иванов и Фербер, 2014. − 352 с.

[14]. Бранцевич, П.Ю. Компьютерные вибродиагностические системы / П.Ю. Бранцевич, С.Ф. Костюк, Е.Н. Базылев, В.Э. Базаревский // Междунар. науч.-техн. конф., приуроченная к 50-летию МРТИ–БГУИР: материалы конф. – Минск : БГУИР, 2014. – Ч. 1, – с. 430–431 .

[15]. Бранцевич, П.Ю. Большие данные в системах вибрационного контроля, мониторинга, диагностики / П.Ю. Бранцевич, Е.Н. Базылев // Неразрушающий контроль и диагностика. − 2016. − № 3. − с. 28-41.

[16]. Бранцевич, П.Ю. Измерительно-вычислительная система распределенного сбора и централизованной обработки виброметрических данных / П.Ю. Бранцевич // Датчики и преобразователи информации систем измерения, контроля и управления. Сборник материалов 12-ой научно-технической конференции с участием зарубежных специалистов. Под ред. В.Н. Азарова. М.: МГИЭМ, 2000. - с. 170-171.

[17]. Базылев, Е. Н. Особенности применения встроенных систем в системах вибрационного контроля, мониторинга, диагностики / Е. Н. Базылев, П. Ю. Бранцевич, С. Ф. Костюк // Международный конгресс по информатике: информационные системы и технологии = International Congress on Computer Science: Information Systems and Technologies[Электронный ресурс] : материалы междунар. науч. конгресса, Республика Беларусь, Минск, 24-27 окт. 2016 г. - Минск : БГУ, 2016. - С. 759-762.

АЛГОРИТМЫ АНАЛИЗА ТОНАЛЬНОСТИ ТЕКСТА

Н.С. Иванин
Студент кафедры информатики БГУИР

А.И. Гербик
Студент кафедры информатики БГУИР

Е.А. Макович
Студент кафедры информатики БГУИР

М.В. Аксамит
Студентка кафедры информатики БГУИР

П.Е. Дорошкевич
Студент кафедры информатики БГУИР

А. И. Свито
Студент кафедры информатики БГУИР

Белорусский государственный университет информатики и радиоэлектроники, Республика Беларусь
E-mail: nikivnik@gmail.com, alexander.gerbik@gmail.com, egormakovich@rambler.ru, rikka1128@gmail.com, pavel.darashkevich@gmail.com, alexandervirk@gmail.com

Abstract. Sentiment Analysis (SA) is an ongoing field of research in text mining field. SA is the computational treatment of opinions, sentiments and subjectivity of text. In this paper various algorithms for sentiment analysis are studied and challenges and applications appear in this field are discussed.

Рост информатизации общества и проникновение технологий во все сферы деятельности человека повлек за собой накопление больших объемов данных, в частности текстовых. Текстовая информация играет важную роль в деятельности человека, так как это наиболее распространенный и универсальный способ представления знаний человека об окружающем мире. Поэтому в настоящее время является актуальной задача обработки текстовых данных. Ручная обработка невозможна из-за большого объема накопленных данных, автоматическая обработка осложняется отсутствием структурированности в текстовых данных, неоднозначностью трактовки значений слов, наличием многочисленных исключений из правил естественного языка и т.д.

Задача анализа тональности текста (Sentiment analysis) является одной из задач обработки естественного языка (Natural Language Processing). Целью анализа тональности является нахождение мнений в тексте и определение позиции автора относительно упомянутой темы. Позиция автора может быть различной, и тональная оценка может принимать различные значения. Например: “положительная”, “отрицательная” и “нейтральная” либо “положительная” и “отрицательная”. Данную задачу можно рассматривать как задачу классификации на три и два класса соответственно, далее мы будем рассматривать задачу с двумя возможными вариантами тональной оценки, так как задача классификации на три и более класса является более

сложной в техническом отношении. Для решения задачи классификации эффективными являются методы машинного обучения с учителем.

Для того, чтобы методы решения задач классификации можно было применить для анализа тональности текста, необходимо текст представить в виде математического вектора. С этой целью применяется векторная модель “мешок слов” - модель текста, предложенная в 1975 году Дж. Солтоном, и в настоящее время одной из самых распространенных в различных областях лингвистических исследований. Текст в данной векторной модели рассматривается как неупорядоченное множество слов. Вектор, являющийся модельным представлением текста в векторном пространстве, образуется упорядочением весов всех слов (включая те, которых нет в конкретном тексте). Размерность этого вектора равна количеству различных слов во всей коллекции, и является одинаковой для всех текстов коллекции.

Так как в естественных языках встречаются устойчивые словосочетания, смысл которых отличается от смысла входящих в их слов (например, “give up” в английском языке) целесообразно применять модель не “мешок слов”, а “мешок N-грамм”. N-грамма - последовательность из N элементов (слов). Последовательность из трех элементов называют триграмма, последовательность из двух элементов называется биграмма. N-граммы меньшей длины, как правило, дают лучшие результаты, чем N-граммы большей длины, т.к. обучающая выборка в большинстве случаев недостаточно большая для нахождения статистических закономерностей N-грамм большой длины. Для многих практических приложений можно получить хороший результат, используя в качестве признаков одиночные слова и биграммы.

Модель “мешок слов” некорректно работает со словами, меняющими тональность выражения на противоположное. Например, фразы “мне нравится этот фильм” и “мне не нравится этот фильм” будут иметь положительную тональность, хотя у второй фразы она должна быть отрицательной. Чтобы решить эту проблему, можно объединять слово “не” со следующим словом, в результате в данном примере мы получим слово “не-нравится” и модель будет работать корректно. Также эту проблему можно решать при помощи N-грамм, но, как правило, это вынуждает использовать N-граммы большей длины.

Для ликвидации неоднозначности, вызванной возможностью одного и того же слова быть различными частями речи, применяется тегирование частей речи - определение для каждого слова в предложении его части речи по положению в предложении и/или грамматической форме.

Полученную задачу классификации можно решить различными методами машинного обучения: наивный байесовский классификатор, логистическая регрессия, метод опорных векторов, методы нейронных сетей и т.д. Сравнив их временную сложность, качество полученных моделей, масштабируемость можно выбрать наиболее подходящий для конкретных данных и конкретной задачи.

Алгоритмы анализа тональности текстов предназначены для определения тональности целого текста либо его фрагмента. В таком подходе предполагается, что исходный текст является мнением автора о каком-то одном конкретном объекте, например, ресторане или книге. Однако в некоторых доменах в отзыве о объекте или сущности так же содержится мнение автора о ее составляющих. Например, в отзывах о мобильных телефонах могут быть оценены такие части телефона, как экран, камера, аккумулятор, батарея. Поэтому после решения задачи анализа тональности текстов начали разрабатываться алгоритмы для анализа мнений по конкретным свойствам или частям (такие свойства или части называются аспектами). О таких свойствах или частях автор отзыва может высказывать мнения с различной тональностью и основной задачей алгоритмов анализа тональности является выделение аспектов и определение тональности отзыва пользователя об каждом аспекте.

Согласно [1] мнение (или регулярное мнение) это набор из пяти элементов $(e_i, a_{ij}, oo_{ijkl}, h_k, t_l)$, где e_i – это имя сущности; a_{ij} – это один из аспектов сущности; oo_{ijkl} – это тональность мнения автора о аспекте a_{ij}, относящемуся к сущности e_i; h_k – автор мнения;

t_l – время, когда автор h_k высказал свое мнение. Регулярное мнение – это позитивное или негативное настроение, отношение, эмоция об объекте или аспекте объекта, которое высказал автор. Тональность мнения oo_{ijkl} может быть положительной, отрицательной или нейтральной, либо же измеряться в некотором интервале, например, от 0 до 1.

Довольно часто в отзывах пользователей можно встретить мнение об объекте в целом, например, «отличный смартфон». В [2] делается предположение о том, что такую категорию можно рассматривать как аспектную. Также существует возможность объединить аспекты в аспектные категории. Для случая смартфонов такие аспекты как разрешение, цветопередача, диагональ могут быть объединены в аспектную категорию дисплей.

Когда человек высказывает свое мнение о чем-либо, то его высказывание имеет некоторую цель. Такой целью служит аспект или тема, которые в дальнейшем будут извлекаться из высказывания. Таким образом, основой задачи извлечения аспектов является определение оборотов, характеризующих отношение автора, и аспекта, к которому автор высказывает свое отношение. Обороты, выражающие настроение, могут выполнять две функции: показывать положительное или отрицательное отношение и быть неявным аспектом, например «этот телефон большой», «большой» это прилагательное, характеризующее отношение автора, но также это неявный аспект размер. Как правило, в качестве аспектов выступают существительные и именные группы[3]. Длина извлекаемых именных групп при этом обычно не превосходит трех.

В [4] выделяют 4 основных подхода к извлечению явных аспектов:

– извлечение на основе часто встречающихся существительных и именных групп;

– извлечение на основе отношений между оценочными оборотами и аспектами;

– извлечение на основе машинного обучения с учителем;

– извлечение на основе статистических тематических моделях.

В подходе извлечения аспектов на основе часто встречающихся существительных и именных групп осуществляется поиск явных оценочных оборотов, как было отмечено выше, это существительные и именные группы. Они извлекаются из большого числа отзывов из определенной области.

В работе [3] для извлечения аспектов используется алгоритм, основанный на CBA[5]. Перед началом работы к каждому отзыву необходимо осуществить предобработку. Это необходимо для исключения слов, которые обычно не являются аспектами. Предобработка включает удаление стоп-слов, стемминг, лематизацию и исправление написания слов. На следующем шаге алгоритм CBA извлекает часто встречающиеся множества элементов. Каждый элемент в этом множестве это возможный аспект. Для извлечения полезных и подлинных аспектов используется фильтрация. Авторы предлагают использовать два типа фильтрации: фильтрация на основе компактности (среди кандидатов длины 2 и более удаляются те, составляющие которых отстоят друг от друга на большом расстоянии) и фильтрация лишних кандидатов (среди кандидатов длины 1 удаляются те, которые определенное число раз входят в кандидаты большей длины). Затем осуществляется поиск оценочных оборотов. Для каждого отзыва, который содержит аспект, извлекается ближайшее прилагательное. Если такое прилагательное найдено, то оно рассматривается как оценочный оборот. Так же данный подход позволяет извлекать аспекты, упомянутые только несколькими пользователями. Для этого из каждого отзыва, который не содержит аспектов, но содержит оценочный оборот, извлекается наиболее близкое к оценочному обороту существительное или именная группа.

В [6] предлагается система, построенная на основе домен независимой системы извлечения информации KnowItAll[7]. На первом этапе своей работы предложенная система извлекает существительные и именные группы из отзывов, оставляя при этом только те, частота встречаемости которых больше определенного уровня. После этого каждой именной группе присваивается оценка с помощью оценивающего модуля. Оценка выставляется на основе вычисления PMI[8]. Система использует явные аспекты для извлечения оценочных оборотов.

Если в предложении содержится аспект, то она использует определенные шаблоны извлечения оценочных оборотов. В системе, описанной выше, использовалась похожая идея, однако в данной системе для извлечения применяется парсер, генерирующий синтаксические зависимости.

Как было описано выше, оценочные обороты имеют цель. Довольно часто найти оценочные обороты не является сложной задачей, поэтому для извлечения аспектов достаточно найти цели. На этой идее основан метод извлечения аспектов на основе отношений между оценочными оборотами и аспектами.

В работе [9] для извлечения аспектов используется синтаксический анализатор, который генерирует граф грамматической зависимости. Этот графи используется для получения зависимостей между аспектами и направленными на них оценочными оборотами. В этой системе применяется Stanford Parser (http://www-nlp.stanford.edu/software/lex-parser.shtml). Этот парсер используется для определения наиболее короткого расстояния от аспекта до оценочного оборота. Затем производится стемминг и частеречная разметка. После этого извлекается размеченная часть между аспектом и оценочным оборотом, например, в предложении «This smartphone is great» smartphone является аспектом, а great оценочным оборотом. Это предложение будет размечено следующим образом «smartphone(NN) – nsubj – is(VBZ) – dobj – great (JJ)». После удаления встречаемых редко шаблонов оставшиеся шаблоны используются как шаблоны отношений между аспектами и оценочными оборотами для извлечения аспектов.

В последнее время довольно активное распространение получили алгоритмы машинного обучения. Они находят активное применение в задаче извлечения информации. Извлечение аспектов из отзывов так же относится к этой задаче, что дает возможность применения алгоритмов машинного обучения для извлечения аспектов. Существует два подхода при использовании алгоритмов машинного обучения в нашей задаче: методы, использующие для обучения заранее подготовленный список аспектов и методы, основанные на разметке последовательности слов. Наибольшее распространение получили методы на основе скрытых марковских моделей и условных случайных полей.

В работе [10] были для извлечения аспектов были применены условные случайные поля. Аспекты извлекались из предложений, содержащих оценочные обороты. На вход модели условного случайного поля были переданы следующие параметры:

– токен - этот параметр приставляет собой текущий токен;

– часть речи – этот параметр представляет собой часть речи текущего токена. Так же этот параметр может служить для разрешения лексической неоднозначности;

– путь зависимости – путь, получаемый в синтаксическом дереве между аспектом и оценочным оборотом. Для получения пути зависимости используется Stanford Parser (http://www-nlp.stanford.edu/software/lex-parser.shtml);

– расстояние между словами.

В этой системе возможные метки использовались в соответствии со схемой Inside-Outside-Begin: метка B-target означала начала аспекта, I-target означала продолжение аспекта, а метка O использовалась для обозначения других токенов.

Статистические тематические модели используются для определения тем на основе большой коллекции документов. Тематическое моделирование относится к обучению без учителя, который считает, что текст состоит из некоторого числа тем, а темы являются вероятностным распределением слов. Тематические модели могут быть применены для извлечения аспектов, если каждый аспект будет рассматриваться как униграммная языковая модель [11].

В работе [12] предложена модель, которая является смесью моделей для аспектного анализа и анализа тональности. Это модель состоит из аспектной модели, модели анализа положительной тональности и модели анализа отрицательной тональности. Эти модели были обучены на некоторых тренировочных тестовых данных. Предложенная модель базируется на pLSA [13].

Подходы, основанные на машинном обучении, показывают неплохие результаты. Однако они требуют для обучения размеченные данные, а этот процесс довольно трудоемкий, к тому же полученные модели являются доменно-зависимыми. Подходы, основанные на часто встречающихся существительных и именных группах и на отношениях между оценочными оборотами и аспектами позволяют избежать этих проблем, однако они показывают меньшую точность при работе.

В дальнейшем предполагается реализовать рассмотренные методы, а также рассмотреть возможность применения методов машинного обучения — регрессионного и структурированного вариантов SVM, Gradient boosting.

Литература

[1]. Liu B., Zhang L. A survey of opinion mining and sentiment analysis // Mining Text Data. Springer: US, 2012. P. 415-463.

[2]. Лукашевич Н. В. Автоматический анализ тональности текстов по отношению к заданному объекту и его характеристикам // Russian Digital Libraries Journal. — 2015. — Т. 18, № 3-4 . — С. 88–119.

[3]. Hu M., Liu B., Mining opinion features in customer reviews, Proceedings of the 19th national conference on Artifical intelligence, p.755-760, July 25-29, 2004, San Jose, California.

[4]. B. Liu. Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies, pages 1--167, 2012.

[5]. Liu, B., Hsu, W., Ma, Y. 1998. Integrating Classification and Association Rule Mining. KDD-98, 1998.

[6]. Popescu A., Extracting product features and opinions from reviews // Natural language processing and text mining. A. Popescu et al. Springer: London. 2007. P. 9-28.

[7]. Etzioni O., Unsupervised named-entity extraction from the Web: An experimental study, Artificial Intelligence, O.Etzioni, [et al], v.165 n.1, p.91-134, June 2005 .

[8]. Turney P.D., Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL, Proceedings of the 12th European Conference on Machine Learning, p.491-502, September 05-07, 2001.

[9]. Zhuang L., Jing F., Zhu X. Movie review mining and summarization // Pro-ceedings of ACM International Conference on Information and Knowledge Management (CIKM-2006), 2006. P. 43-50.

[10]. Jakob N, Gurevych I., Extracting opinion targets in a single- and cross-domain setting with conditional random fields, Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, p.1035-1045, October 09-11, 2010, Cambridge, Massachusetts.

[11]. Chu W.W. Data Mining and Knowledge Discovery for Big Data: Methodologies, Challenge and Opportunities (Studies in Big Data, Springer. 2013.

[12]. Mei Q., Ling X., Wondra M., Su H, Zhai C, Topic sentiment mixture: modeling facets and opinions in weblogs, Proceedings of the 16th international conference on World Wide Web, May 08-12, 2007, Banff, Alberta, Canada.

[13]. Hofmann T., Probabilistic latent semantic analysis, Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence, p.289-296, July 30-August 01, 1999, Stockholm, Sweden.

РЕАЛИЗАЦИЯ МУЛЬТИМЕДИЙНОГО IOT РЕШЕНИЯ С ИСПОЛЬЗОВАНИЕ ОБЛАЧНЫХ ТЕХНОЛОГИЙ

Е.Н. Побыванец
Инженер-программист ЗАО «Итранзишэн»,
магистрантом кафедры информатики БГУИР

Белорусский государственный университет информатики и радиоэлектроники, ЗАО «Итранзишэн», Республика Беларусь
E-mail: e.pobyvanets@itransition.com

Abstract. Main goal of this research is to describe possible implementation of multimedia IoT solution, which will allow user to play multimedia content, such as music or video, on remotely controlled devices. Meanwhile, played multimedia content should adapt to user preferences and provide recommendations for end-user. During this research, I was able to determine most possibilities and limitations of such systems, describe an architecture of application and start working on prototype solution, based on cloud technologies.

На сегодняшний день «умных» устройств, подключенных к сети, превышает количество людей на планете Земля. Шутка ли это? Вовсе нет, по статистике, на 2016 год к сети подключено более 8 миллиардов устройств [1]. В связи с таким количеством подключенных устройств не удивительно, что все чаще и чаще в сети проскакивает термин Internet of Things (IoT, Интернет вещей). Что же такое Интернет вещей? Согласно IT глоссарию Интернет вещей – это сеть, состоящая из физических объектов, которые содержат встроенные технологии, позволяющие обмениваться состоянием этих объектов с окружающими объектами и устройствами [2].

Первое упоминание об Интернете вещей датируется 1990 годом, когда широкой общественности был продемонстрирован тостер, который имел возможность подключения к интернету, но широкое развитие термин получил только начиная с 2008 года. С тех пор, мир уже успел увидеть и удивиться многим устройствам, и даже привыкнуть к ним. Умные браслеты, умные часы, сервис доставки посылок дронами от Amazon, Google Home – все эти названия уже на слуху и являются частью нашей повседневной жизни.

«Умные вещи» прочно закрепились во многих областях, и, по прогнозам аналитиков, предполагается, что они станут активными участниками социальных, информационных и бизнес-процессов. Сейчас же интернет вещей служит для связи более мелких, казалось бы, несовместимых процессов в одну большую, единую сеть (рисунок 1).

Преимущества IoT-решений достаточно очевидны – возможность автоматизировать многие процессы в нашей жизни, позволить умным алгоритмам избавить нас от рутины, или же наоборот, наиболее быстрым способом получить уведомление о том, что человек мог бы и не заметить. Но у этих же решений есть и недостатки. Каждое такое решение требует автономности каждого устройства, позволяющее ему работать практически без перерывов. Также возникает проблема стандартизации – каждый разработчик IoT-решения пытается адаптировать его под себя, поскольку на данный момент нет общепринятых стандартов, позволяющих интегрировать все IoT устройства друг с другом [3].

В данной работе рассматривается возможность создания мультимедийного IoT решения,

представленного в виде системы взаимодействующих друг с другом приложений, которая позволяла бы в «умном» режиме (учитывая предыдущие воспроизведения, а также текущую громкость воспроизведения, время суток и другие условия) воспроизводить мультимедийный контент с помощью смарфтона, или любого другого устройства, имеющего доступ в сеть интернет. Исходя из сформулированной цели, были описаны следующие требования к системе.

1 Система должна иметь возможность принимать управляющие команды от клиентского приложения.

2 Система должна иметь принимающее устройство, способное воспроизводить мультимедийный контент на доступных устройствах.

3 Система должна иметь возможность сохранять информацию о воспроизведенном мультимедийном контенте и другую информацию о воспроизведении, уметь анализировать эту информацию и на ее основе анализа иметь возможность предоставить рекомендации к контенту.

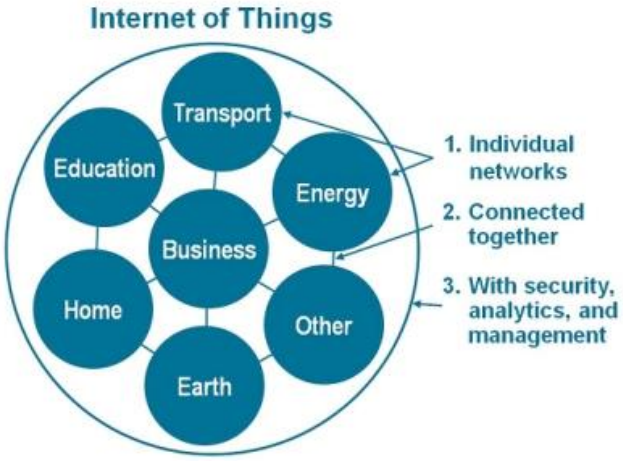


Рис. 1. Сферы, связанные с помощью Интернета вещей

Основываясь на цели и сформулированных требованиях были выделены следующие задачи для реализации данного решения:

4 Разработать архитектуру приложения, позволяющего поддерживать взаимодействие «Клиент – Сервер – Управляющее устройство».

5 Проанализировать возможности использования различных устройств в качестве компонент системы.

6 На основе информации, представленной в метаданных мультимедийного контента, проанализировать возможные рекомендационные системы и найти наиболее подходящую.

7 Разработать соответствующие компоненты системы.

Архитектура системы приведена на рисунке 2.

Анализируя требования к системе, необходимо было найти подходящее решение, которое могло бы помочь в унификации запросов к принимающему устройству. Таким решением может быть RESTful Server – сервер, принимающий запросы в формате REST (Representative Stateless Transfer). REST запрос содержит в себе данные, необходимые для передачи на сервер и HTTP-глагол, позволяющий определить тип операции. Такая структура запроса позволяет отправлять унифицированные запросы с любого устройства, без привязки к особенностям системы. Для сохранения информации об уже обработанном мультимедийном контенте, RESTful Server имеет доступ к базе данных MongoDB. MongoDB – NoSQL база данных, позволяющая быстро работать с не реляционными данными) и обеспечивающая достаточно про-

стую интеграцию с множеством сервисов и библиотек. Анализируя метаданные мультимедийного контента, можно заметить, что эти данные являются не реляционными, что и обосновывает выбор этой базы данных.

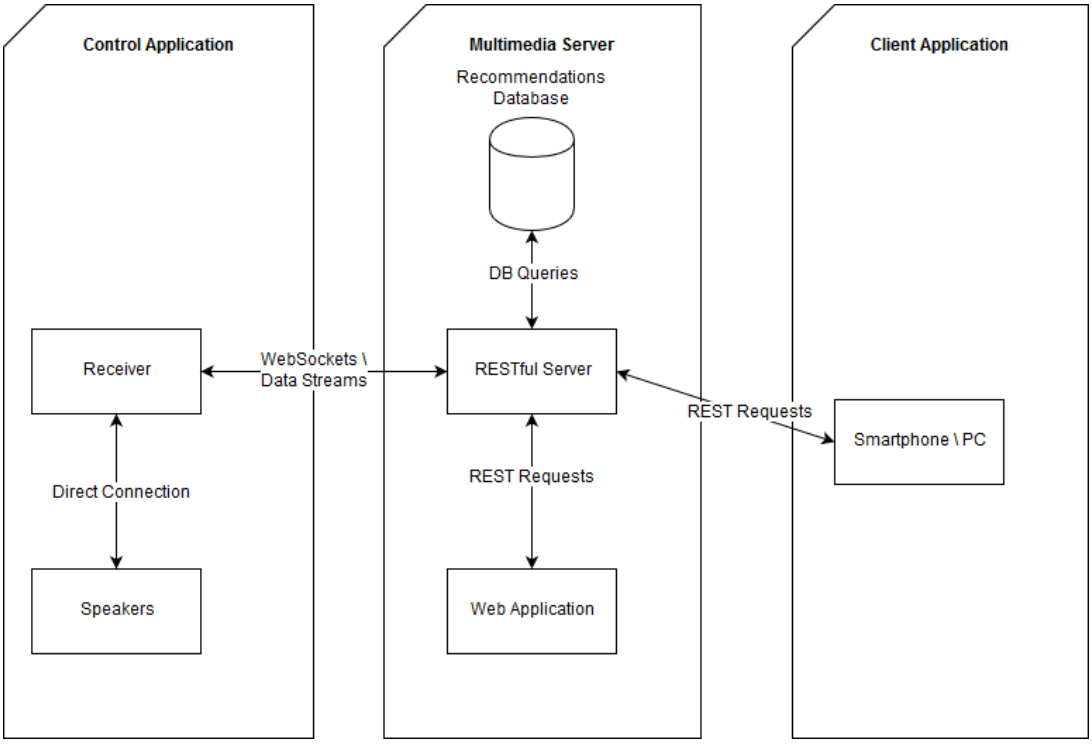


Рис. 2. Архитектура системы IoT-решения

В качестве принимающего устройства может быть использовано любое устройство, поддерживающее подключение к мультимедийному устройству, а также способному принимать информацию из сети. Наиболее подходящим по данным критериям, а также наиболее простым для прототипирования устройством является смартфон на платформе Android, поскольку большинство смартфонов имеют готовый интерфейс для коммуникации как с воспроизводящими устройствами, так и с сервером, с помощью приложений. Также подходящим устройством может являться выделенный сервер \ домашний компьютер, но общая конструкция может получится достаточно громоздкой. Коммуникация между смартфоном и мультимедийным сервером будет осуществляться с помощью двух технологий. Первая – WebSocket – позволяет осуществлять обмен информацией в режиме реального времени, что положительно сказывается на времени отклика устройства. Эта технология позволяет передавать простые команды, такие как «Пауза», «Начать воспроизведение» и т.д. Вторая же технология – Data Streams – позволяет осуществлять стриминг мультимедийного контента, без ожидания его предварительной загрузки. Такое решение было принято в связи с ограничением на доступную память.

Для контроля над текущими операциями, а также для резервного доступа к системе, в качестве надстройки над RESTful Server'ом предлагается реализовать простое клиентское веб-приложение, предоставляющий простой функционал по контролю над системой.

Отдельного рассмотрения требует рекомендационный сервис, расположенный на мультимедийном сервере. В качестве ядра, рекомендационный сервис использует Open Source библиотеку Apache Mahout – библиотека, предназначенная для машинного обучения и имеющая в своей реализации следующие группы алгоритмов, применимые к данному решению [4]:

1 Алгоритмы рекомендательной системы.

Apache Mahout предоставляет возможность использовать алгоритмы, позволяющие осуществлять построение рекомендательной системы на основе коллаборативной фильтрации,

что позволяет достаточно быстро построить обученную модель на основе большого объема данных о предпочтениях пользователей в сфере мультимедиа [5].

8 Алгоритмы классификации.

В дополнении к алгоритмам, связанных с рекомендательными системами, Apache Mahout предоставляет возможность использовать алгоритмы классификации мультимедийного контента по метаданным, что позволяет в конечном итоге, уточнить рекомендательную систему и построить обучаемую модель, на основе правильной классификации нового контента.

Последнее, но не менее важное, что стоит рассмотреть в данном исследовании, это вопрос информационной безопасности, связанный с передачей данных. В связи с возможностью управления устройствами из сети, необходимо обеспечить механизм, позволяющий однозначно идентифицировать пользователя, имеющего доступ к данной системе. Для решения этой проблемы предлагается использовать Token Based Authentication, и в частности bearer-token. Механизм аутентификации через токены прекрасно себе зарекомендовал при совмещении в единую систему различных устройств и приложений [6], т.к. он предоставляет возможность однозначно идентифицировать пользователя по совокупности введенных им данных и в дальнейшем использовать сгенерированный на определенное время токен, как подтверждение того, что пользователь имеет право на осуществление операции.

Подводя итоги исследования, хотелось бы отметить, что сфера IoT имеет очень большие перспективы, т.к. все больше крупных компаний выпускают серьезные решения, значительно упрощающие решение повседневных задач.

В данной работе были рассмотрены различные технологии и средства, позволяющие создать мультимедийное IoT решение. На основе требований к системе и поставленных задач была разработана архитектура решения. Анализируя требования к задаче и полученное архитектурное решение, были получены следующие результаты:

Разработана система, состоящая из трех компонентов (клиентское приложение, мультимедийный сервер, управляющие приложение) в которой:

В качестве интерфейса для отправки и обработки команд в сети интернет наиболее подходящей оказалась концепция REST и соответствующий RESTful Server, в связи с широкой степенью унификации данного интерфейса.

В качестве принимающего устройства, может быть использовано любое устройство, поддерживающее возможность подключение к мультимедийному интерфейсу и сети интернет, но наиболее подходящим для быстрой разработки является смартфон на базе Android.

В качестве ядра для создания системы рекомендации была выбрана библиотека с открытым исходным кодом Apache Mahout, имеющую широкий набор реализованных алгоритмов.

Литература

[1]. FABS in the Internet of Thins Era (2013), David Lammers, https://www.eiseverywhere.com/file_uploads/27ceb1798b372d92a7fd66726e007473_Applied-2.pdf

[2]. Gartner IT Glossary (2017), Gartner Inc, http://www.gartner.com/it-glossary/internet-of-things

[3]. «Интернет вещей – а что это?» (2014), Николай Пилипенко, https://geektimes.ru/post/149593/

[4]. «Mahout in Action» (2011), Sean Owen, Robin Anil. ISBN 9781935182689, 416 стр.

[5]. Million Song Dataset (2012), Laboratory for the Recognition and Organization of Speech and Audio, https://labrosa.ee.columbia.edu/millionsong/

Token Based Authentication Made Easy (2016), https://auth0.com/learn/token-based-authentication-made-easy

BIG DATA. ТРАНСФОРМАЦИЯ МАГИСТЕРСКИХ ПРОГРАММ

И.И. Пилецкий
Научный руководитель совместной лаборатории БГУИР-IBA и АЦКТ IBM, архитектор отделения по программному обеспечению IBA IT Park, доцент кафедры информатики БГУИР, кандидат физико-математических наук

А.Е. Лещев
Заместитель декана по учебной работе факультета компьютерных систем и сетей БГУИР, магистр технических наук

В.И. Козуб
Ассистент кафедры информатики БГУИР

Белорусский государственный университет информатики и радиоэлектроники, Республика Беларусь
E-mail: ianmenski@gmail.com, leschov@bsuir.by, kozub@bsuir.by

Abstract. This paper describes difficulties which we've experienced with training of IT professionals in Big Data field. IT skills with top demand on market are also given. Master degree study programs for speciality 'Big Data Processing' are considered as well as difficulties with training of master degree students in this field. These difficulties are caused by flexible and rapidly changing requirements for this speciality which can't be fulfilled by university itself. Enterprise technologies and solutions can be used in education as a tool for developing necessary skills. Using of these technologies in master degree study programs is also considered.

1. Принципы трансформации. В последние годы мир IT (и не только он) быстро изменяется. Данными является всё то, что мы видим, слышим, что получаем от различных приборов и датчиков. Это огромное количество данных, которые плохо структурированы, например, текстовые данные из Интернет-источников, фотографии, видеозаписи, электронные журналы, геопространственные данные и др.

Так, 90 % данных в мире было создано в последние два года [1-6]:

– каждый день по всему миру продается 2 млн. мобильных устройств;

– по данным GSMA, в 2015 году зафиксировано 7,23 млрд. мобильных соединений;

– в 2016 году более 70 % клиентов зависят от социальных сайтов (от советов до покупки);

– к 2020 году общая рабочая загрузка облачных ЦОД будет приближаться к 90 %, роль традиционных ЦОД будет сокращаться до 8 % (в 2016 году данное соотношение составляло 81 % и 19 % соответственно);

– к 2020 году будет 50 миллиардов IoT-устройств, подключённых к Интернету;

– продолжается рост API экономики, глобальные продажи (E-commerce, M-commerce) неуклонно растут и к 2020 году превысят 600 миллиардов долларов.

По данным корпорации IBM, наблюдается значительный рост использования общественных сетей. Наступает новая постиндустриальная эра, в которой IT займут положение ключевой отрасли в экономике, подобно тому как ткацкое производство занимало эту роль в XVII-XVIII веках, а машиностроение – в XIX-XX веках.

Наиболее востребованными навыками (специальностями) в IT во всём мире в 2016 году по данным LinkedIn являлись следующие (см. таблицу 1). Для получения данных было проанализировано более 330 миллионов LinkedIn профайлов [7-8].

Таблица 1. Наиболее востребованные навыки в IT в 2016 году

#1 Cloud and Distributed Computing	Облачные вычисления и распределённые вычисления
#2 Statistical Analysis and Data Mining	Статистический анализ и интеллектуальный анализ данных
#3 Web Architecture and Development Framework	Архитектура и разработка веб-сайтов
#4 Middleware and Integration Software	Промежуточное и интеграционное ПО
#5 User Interface Design	Дизайн пользовательского интерфейса
#6 Network and Information Security	Сетевая и информационная безопасность
#7 Mobile Development	Разработка ПО для мобильных устройств
#8 Data Presentation	Представление данных (визуализация)
#9 SEO/SEM Marketing	SEO/SEM маркетинг
#10 Storage Systems and Management	Системы хранения данных и управление ими

В табл. 2 приведена тенденция к требуемым специальностям на протяжении 2015 и 2016 годов.

Таблица 2. Тенденция к требуемым специальностям на протяжении 2015 и 2016 годов

The hottest Skills of 2015 on LinkedIn	Top Skills of 2016 on LinkedIn
Cloud and Distributed Computing	Cloud and Distributed Computing
Statistical Analysis and Data Mining	Statistical Analysis and Data Mining
Marketing Campaign Management	Web Architecture and Development Framework
SEO/SEM Marketing	Middleware and Integration Software
Middleware and Integration Software	User Interface Design
Mobile Development	Network and Information Security
Network and Information Security	Mobile Development
Storage Systems and Management	Data Presentation

В IT-сфере появилась новая проблема: адаптация и трансформация образовательных программ к современным требованиям. Основные принципы трансформации программ:

– программы должны отражать тенденции в IT, содержать современный контент;

– помогать студентам в карьерном плане;

– предоставлять новые подходы к образованию, связанные с новыми технологическими возможностями по обучению.

2. База для трансформации. Первое направление (на кафедре информатики БГУИР факультета КСиС) связано с *обучением студентов в совместной лаборатории БГУИР-IBA* для работы с Big Data. Обучение проводится на облачной платформе ЦОД БГУИР [9-10] и платформе IBM Bluemix [11].

Платформа ЦОД БГУИР обеспечивает сервисные модели IaaS и PaaS. Основу данной платформы составляют два семейства продуктов: IBM InfoSphere (InfoSphere BigInsights и InfoSphere Streams) и Open Source решения: Apache Storm, Apache Spark и др.

Новейшие облачные технологии IBM и сервисные модели типа PaaS и SaaS платформы IBM Bluemix [11] использовались как базис для выполнения практических работ студентами.

Программы для обучения в совместной лаборатории строятся на основе регулярно уточняющихся курсов IBM и материалов, подготовленных преподавателями. Так, на протяжении последних трёх учебных лет студенты факультета КСиС проходят в лаборатории начальный курс IBM по Big Data и получают сертификаты IBM [9].

Второе, основное направление, связано с *обучением студентов в магистратуре* кафедры информатики по специальности «Обработка больших объёмов информации».

В 2015 году сотрудниками БГУИР был разработан образовательный стандарт для второй

ступени образования (магистратуры) ОСВО 1-40 81 04-2015 «Обработка больших объемов информации», который был утвержден Министерством образования Республики Беларусь. Стандарт обязателен для применения во всех учреждениях высшего образования Республики Беларусь, реализующих образовательные программы магистратуры.

В стандарте предусмотрены *дисциплины государственных компонент:* «Технологическая платформа по управлению большими данными», «Архитектурные решения для обработки больших объемов информации», «Модели и методы обработки и анализа больших объёмов информации».

В соответствии с данным стандартом в 2016 году разработаны три учебные программы БГУИР: «Технологическая платформа по управлению большими данными», «Архитектурные решения для обработки больших объемов информации», «Модели и методы обработки и анализа больших объёмов информации» – по которым в 2016 году начата подготовка магистров по специальности 1-40 81 04 Обработка больших объемов информации».

Общая тематика этих программ позволяет охватить большинство аспектов «Обработки больших объемов данных». При проведении занятий были рассмотрены такие темы, как: что такое «большие данные», источники больших данных, методы доступа к ним, проблема обработки больших объёмов данных, вычислительные модели многопоточной параллельной обработки, файловые системы (HDFS, GPFS, RDD), NoSQL базы данных (типа Riak, Amazon Dynamo, Redis, HBase, CouchDB, MongoDB, Infinite Graph, Neo4J), методы доступа и хранения данных (плоские файлы, JSON, CSV, XML и NoSQL базы данных), платформы, применяемые для работы с большими данными (Open Source, Enterprise Solutions, Cloud Based), обработка данных «в покое», обработка потока данных, архитектурные решения систем вертикального и горизонтального масштабирования, платформы с открытым кодом и реализующие промышленные решения, сравнительный анализ архитектурных решений платформ для обработки больших объёмов данных, системы для обработки больших объёмов данных, языки параллельного программирования для анализа структурированных и неструктурированных данных больших объёмов (Pig, Jaql, Hive, SQL, R), методы, средства и аналитические алгоритмы анализа структурированных и неструктурированных данных, описательная, прогнозная и директивная аналитика, методы нахождения паттернов и аномалий, модели и методы визуального представления данных, примеры графической интерпретации больших объёмов информации.

Третье, *научно-исследовательское направление*, связано с исследованиями, которые проводятся при написании диссертаций в магистратуре и аспирантуре.

Здесь выполняются различные работы, связанные с исследованиями методов и технологий обработки естественного языка из Интернет источников, разработкой и реализацией различных IoT-решений с использованием облачных технологий и когнитивных технологий, исследованием методов и средств анализа мультимедийной информации (в том числе и фото), других методов и алгоритмов машинного обучения, методов и средств аналитических решений (в том числе и в сфере транспорта).

В качестве учебной базы для проведения занятий применялись:

– созданный вычислительный кластер в локальном центре обработки данных БГУИР (ЦОД БГУИР) с ПО для обработки больших объёмов данных [9];

– платформы Apache Hadoop, Cloudera Hadoop Distribution Platform, Apache Spark, Apache Storm;

– базы данных Riak KV, Amazon Dynamo, Redis, CouchDB, MongoDB, Infinite Graph, Neo4J;

– платформы IBM BigInsights и IBM InfoSphere Streams;

– облачная платформа IBM Bluemix;

– когнитивные сервисы IBM Watson.

В настоящее время, наряду с уже существующей базой, для организации и проведения занятий требуются *индустриальные технологии*, так как решить проблему адаптации и трансформа-

ции образовательных программ в сфере IT к современным требованиям индустрии силами университетов невозможно.

Индустриальные технологии в образовании для формирования необходимых навыков должны обеспечивать:

– доступ к технологиям и решениям ведущих мировых лидеров в IT, предоставляющих как Open Source, так и Enterprise решения;
– свободный доступ для университетов;
– использование облачных ресурсов для выполнения практических и лабораторных работ.

В качестве ресурсов могут быть использованы различные (как платные, так и бесплатные) Интернет ресурсы. Например, на сайте IMS GLOBAL Learning Consortium [12] предлагаются разнообразные сертифицированные курсы (IMS Certified Product Directory), LinkedIn на своем сайте [13] предлагает множество курсов для IT-специалистов.

На LinkedIn для каждой специальности предлагаются курсы и открытые рабочие вакансии, так для Cloud and Distributed Computing доступны курсы Cloud Computing, Big Data, Hadoop, Amazon Web Services, а для Statistical Analysis and Data Mining – курсы R, SPSS, Data Analysis. Курсы, как правило, платные.

Мировые лидеры индустрии IT также предлагают различные программы для обучения.

Уже длительное время БГУИР плодотворно сотрудничает с компаниями IBM и IBA Group.

В 2008 году была создана научная учебно-производственная лаборатория «Информационных технологий» кафедры Информатики БГУИР и СП ЗАО «Международный деловой альянс» (IBA Group). Благодаря успешной реализации учебно-исследовательских проектов в 2011 году в БГУИР корпорация IBM отрыла первый в Республике Академический центр компетенции технологий IBM на базе совместной лаборатории БГУИР и группы компаний IBA.

Кафедра Информатики БГУИР и факультет КСиС сотрудничают с корпорацией IBM по программе IBM Academic Initiative, что позволяет получать бесплатную поддержку по учебным материалам и программным средствам [14].

Корпорация IBM предлагает и обеспечивает актуальные индустриальные технологии в образовании для формирования необходимых навыков у обучаемых, что позволяет обеспечить адаптацию и трансформацию учебных материалов (лекционных, практических) для магистерских программ к современным условиям.

Сотрудничество БГУИР с корпорацией IBM в образовательной области позволяет преподавателям и студентам пользоваться бесплатно такими ресурсами, как: IBM Academic Initiative, IBM Bluemix, IBM Watson, IBM Big Data University, Облачный университет IBM Watson RCIS [15].

По программе Academic Initiative компания IBM предлагает разнообразие тематических учебных материалов как в виде статей и книг, так и в виде on-line занятий и видеоматериалов на You Tube, например:

– Case management educator guide;
– Cloud and Big Data educator guide;
– Cognitive computing educator guide;
– DevOps Services educator guide;
– Enterprise computing educator guide;
– IBM Bluemix™ educator guide;
– Internet of Things educator guide;
– MobileFirst educator guide;
– NoSQL DBaaS with IBM Bluemix educator guide;
– Security educator guide;
– Watson Analytics educator guide;
– Watson Services on Bluemix educator guide.

Разнообразие материалов призвано помочь различным категориям пользователей найти

подходящий способ получения знаний.

Программа IBM Academic Initiative предлагает не только курсы для преподавателей и студентов, но также открытые рабочие вакансии и требования к ним.

Программа IBM Big Data University прелагает разнообразные многоуровневые курсы, после успешного изучения которых присваивается электронный бэйдж трёх уровней в зависимости от сложности курса.

Облачный университет IBM Watson RCIS был создан сотрудниками компании IBM Москва с целью поддержки студентов в выполнении ими различных заданий в процессе обучения.

Облачный университет IBM Watson RCIS – это виртуальный ресурс, использующий инструменты компании IBM, а также продукты с открытым кодом, для поддержки студентов вузов в выполнении ими учебных заданий, выпускных работ и исследовательских проектов, содержание которых связано с созданием веб-приложений и сервисов. Участниками являются преподаватели университетов и студенты, добровольно присоединяющиеся к сообществу, созданному в социальной сети

Университет создается усилиями вузов-организаторов, в число которых входят *МГТУ им. Баумана, ВШЭ, БГУИР, Университет «Дубна».*

Основными целями проекта являются:

– создание и обновление университетских курсов по информационным технологиям;

– разработка и совершенствование шаблонов курсов и семинарских занятий, создаваемых на основе социальных инструментов;

– совершенствование и аккумулирование знаний, методик и изобретательских подходов для разработки веб-приложений и облачных сервисов; в том числе методов визуального проектирования;

– быстрое прототипирование коммерческих решений для заказчиков;

– создание смотровой площадки для привлечения потенциальных клиентов и заказчиков проектов, размещение решений на IBM Marketplace.

3. Навыки, специальность, карьера

Сотрудничество с корпорацией IBM позволяет БГУИР:

– внедрить индустриальные технологии в образовании для формирования необходимых навыков у студентов;

– обеспечить адаптацию магистерских программ к изменяющимся требованиям IT;

– предоставить ресурсы для обучения в режиме on-line;

– обеспечить выполнение практических и лабораторных работ с помощью облачных ресурсов

В Республике Беларусь, по данным сайта dev.by, зарегистрировано свыше тысячи IT компаний, в которых открыто свыше трёхсот вакансий. Но эти вакансии мало связаны с современными тенденциями в области Big Data.

По данным сайта Парка высоких технологий, в ПВТ зарегистрировано 165 компаний, которые также нуждаются в IT специалистах.

Наиболее перспективными направлениями в настоящее время и недалеком будущем являются: когнитивные системы, приложения и сервисы, аналитика (Watson) + IoT (Internet of Things) – Интернет вещей как сеть взаимодействия M2M (Machine-to-Machine) позволяет создать безлюдное производство, применение беспилотных машин и роботов в армии и на производстве, «Разумные» дома и города, виртуальные глобальные организации, магазины без товаров, логистические предприятия без подвижного состава и т.д. *(А. Сорокин, IBM Восточная Европа \ Азия)*.

Компания IBM предлагает путь карьерного роста через совершенствование навыков и специализации знаний. Требования для занятия вакансии не простые и довольно высокие, но и предложения тоже высокие.

На сайте LinkedIn можно найти около 1300 предложений о найме на работу для наиболее востребованной специальности *Cloud and Distributed Computing*.

Литература

[1]. Digital-статистика по миру за август 2015 года [Электронный ресурс] / WeAreSocial.sg. – 2015-2017. – Режим доступа: http://www.likeni.ru/events/digital-statistika-po-miru-za-avgust-2015-goda/. – Дата доступа: 20.03.2017.

[2]. Мобильных телефонов – больше, чем людей на планете [Электронный ресурс] / Apps4All. – 2016-2017. – Режим доступа: http://apps4all.ru/post/10-09-14-mobilnyh-telefonov-bolshe-chem-lyudej-na-planete. – Дата доступа: 20.03.2017.

[3]. By 2016 Why 70% of Small Businesses Will Depend On Social Media Tool To Reach New Customers [Электронный ресурс] / Smejoinup. – 2015-2017. – Режим доступа: http://smejoinup.com/blog/by-2016-why-70-of-small-businesses-will-depend-on-social-media-tool-to-reach-new-customers/. – Дата доступа: 20.03.2017.

[4]. Cisco Global Cloud Index: Forecast and Methodology [Электронный ресурс] / Cisco Systems. – 2015-2017. – Режим доступа: http://www.cisco.com/c/dam/en/us/solutions/collateral/service-provider/global-cloud-index-gci/white-paper-c11-738085.pdf. – Дата доступа: 20.03.2017.

[5]. Popular Internet of Things Forecast of 50 Billion Devices by 2020 Is Outdated [Электронный ресурс] / IEEE Spectrum. – 2016-2017. – Режим доступа: http://spectrum.ieee.org/tech-talk/telecom/internet/popular-internet-of-things-forecast-of-50-billion-devices-by-2020-is-outdated. – Дата доступа: 20.03.2017.

[6]. The Rise of M-Commerce: Mobile Shopping Stats & Trends [Электронный ресурс] / Business Insider. – 2016-2017. – Режим доступа: http://www.businessinsider.com/mobile-commerce-shopping-trends-stats-2016-10. – Дата доступа: 20.03.2017.

[7]. LinkedIn Unveils The Top Skills That Can Get You Hired In 2017, Offers Free Courses for a Week [Электронный ресурс] / LinkedIn Official Blog. – 2016-2017. – Режим доступа: https://blog.linkedin.com/2016/10/20/top-skills-2016-week-of-learning-linkedin. – Дата доступа: 20.03.2017.

[8]. TOP skills of 2016. Global list [Электронный ресурс] / LinkedIn Week of Learning. – 2016-2017. – Режим доступа: https://learning.linkedin.com/week-of-learning/top-skills. – Дата доступа: 20.03.2017.

[9]. Пилецкий, И. И. Облачная платформа IBM Bluemix для тренинга по технологиям Big Data / И. И. Пилецкий, А. Е. Лещев, В. Н. Козуб // BIG DATA and Advanced Analytics. Использование BIG DATA для оптимизации бизнеса и информационных технологий : сборник материалов II международной научно-практической конференции; Минск, 15-17 июня 2016 г. / редкол. : М. П. Батура [и др.]. – Минск : БГУИР, 2016. – С. 146-153.

[10]. И.И. Пилецкий и др. Виртуальная ИТ среда БГУИР для исследования Big Data и VCL, с. 21-32, BIG DATA and Predictive Analytics. Использование BIG DATA для оптимизации бизнеса и информационных технологий : сборник материалов междунар. науч.-практ. конф. / редкол. : М.П. Батура [и др.]. – Минск : БГУИР, 2015. – 220 с.

[11]. What is Bluemix [Электронный ресурс] / IBM developerWorks. – 2015-2017. – Режим доступа: https://www.ibm.com/developerworks/cloud/library/cl-bluemixfoundry/. – Дата доступа: 20.03.2017.

[12]. Better Learning From Better Learning Technology [Электронный ресурс] / IMS GLOBAL Learning Consortium. – 2016-2017. – Режим доступа: https://www.imsglobal.org/. – Дата доступа: 20.03.2017.

[13]. Education and Instructional Design [Электронный ресурс] / LinkedIn Learning. – 2016-2017. – Режим доступа: https://www.linkedin.com/learning/topics/education-and-instructional-design. – Дата доступа: 20.03.2017.

[14]. Educator guides [Электронный ресурс] / IBM Academic Initiative. – 2016-2017. – Режим доступа: https://developer.ibm.com/academic/educator-guides/. – Дата доступа: 20.03.2017.

[15]. IBM RCIS Watson Cloud Cognitive University [Электронный ресурс] / IBM Developer Works. – 2016-2017. – Режим доступа: https://www.ibm.com/developerworks/community/groups/service/html/communitystart?communityUuid=bc004137-b64a-4378-ac02-2caf59c56c2a

ОЦЕНКА ЭФФЕКТИВНОСТИ АЛГОРИТМОВ СЕГМЕНТАЦИИ ИЗОБРАЖЕНИЙ ДИСТАНЦИОННОГО ЗОНДИРОВАНИЯ ЗЕМЛИ В УСЛОВИЯХ ОГРАНИЧЕННОГО ВРЕМЕНИ ОБРАБОТКИ

М.В. Козак
Студентка кафедры сетей и устройств телекоммуникаций БГУИР

О.М. Альмияхи
Аспирант кафедры сетей и устройств телекоммуникаций БГУИР

В.Ю. Цветков
Заведующей кафедрой сетей и устройств телекоммуникаций БГУИР, доктор технических наук, доцент

Белорусский государственный университет информатики и радиоэлектроники, Республика Беларусь
E-mail: maryiakazak@gmail.com, vtsvet@bsuir.by, almiahi86@yahoo.by

Abstract. Presented an evaluation for the efficiency of segmentation algorithms for images of earth remote sensing in conditions of limited processing time.

Введение. Сегментация изображений является одной из наиболее сложно реализуемых в реальном времени операций. Известные методы сегментации [1, 2 , 3, 4, 5, 6], основанные на разделении и слиянии областей, выращивании областей, водоразделе, анализе гистограмм, к тому же плохо распараллеливаются и требуют много оперативной памяти. При этом во многих задачах допустимо снижение точности сегментации (потеря деталей, имеющих размер менее заданного) для повышения ее скорости и снижения требований к объему памяти. Целью работы является оценка эффективности алгоритмов сегментации изображений ДЗЗ (дистанционного зондирования Земли) в условиях ограниченного времени обработки.

Данные ДЗЗ, накапливаемые в центрах обработки, имеют значительные объемы, что обусловлено невозможностью их сжатия с потерями для последующей обработки, ростом пространственного и спектрального разрешения съемочной аппаратуры летательных аппаратов. Для повышения эффективности обработки больших объемов данных ДЗЗ (ускорение процедур поиска, совмещения и т.д.) может использоваться их предварительная параметризация, например поиск характерных точек (реперов), с помощью которых возможны последующая идентификация и совмещение изображений. Такие точки располагаются около изломов контурных линий изображений. Контурные линии сами по себе также представляют собой объекты идентификации и могут использоваться как для совмещения фрагментов так и для поиска, классификации и распознавания объектов изображений ДЗЗ. Предварительная сегментация изображений ДЗЗ позволяет выделять на них преимущественно замкнутые контурные линии, что может быть использовано для повышения эффективности поиска, классификации и распознавания объектов. Однако, сегментация является наиболее вычислительно сложной операцией, требующей значительного времени и оперативной памяти для вычислений и хранения сегментированных изображений (время вычислений и емкость памяти растут с увеличением числа сегментов). Кроме того, с ростом пространственного разрешения изображений повышается их детализация, объекты изображений становятся текстурными и, как следствие, возрастает число выделяемых на них сегментов. Для управления числом сегментов в [7] пред-

ложены алгоритмы, использующие прореживание пикселей изображения на основе квадросеток пикселей и предварительное квантование изображений с переменным порогом. Кроме того, для снижения влияния текстурного характера областей на число выделяемых сегментов может использоваться предварительная низкочастотная фильтрация изображений.

В качестве тестовой базы для оценки эффективности алгоритмов сегментации использованы фрагменты изображений ДЗЗ, приведенные на рисунке 1.

В качестве критериев эффективности алгоритмов сегментации рассматриваются время сегментации, число выделенных на изображении сегментов, среднеквадратическая ошибка MSE_R восстановления изображения после сегментации с использованием средних значений яркостей сегментов.

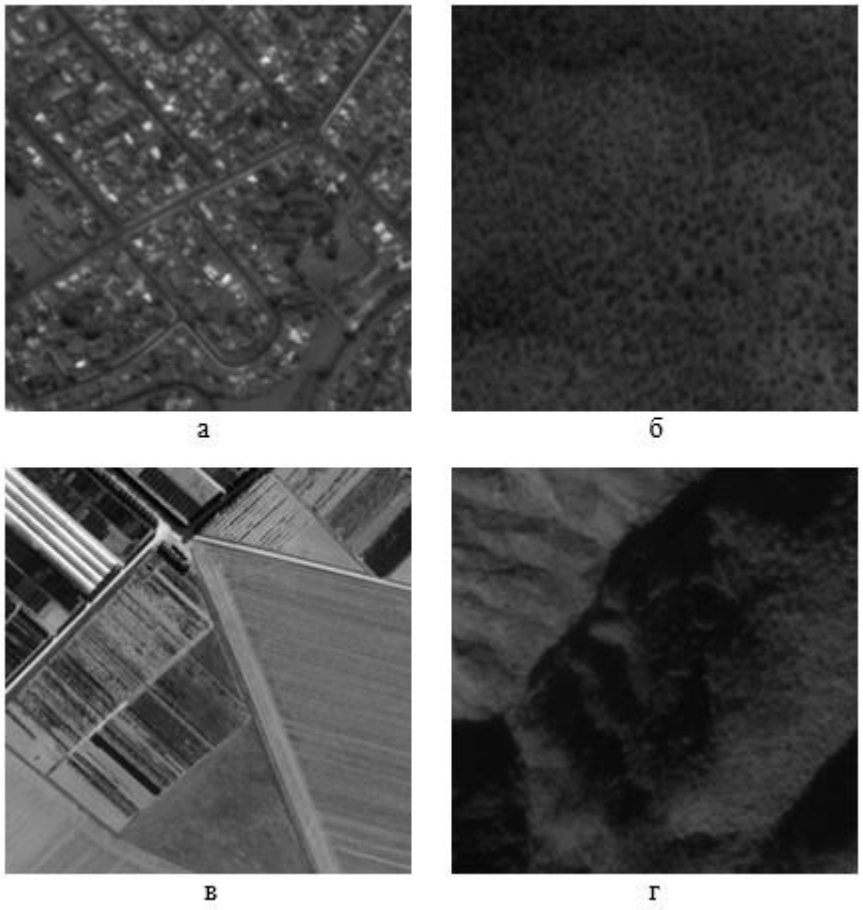


Рис. 1. Тестовые изображения: а – изображение «Город»; б – изображение «Лес»; в – изображение «Степь»; г – изображение «Горы»

Оценка эффективности алгоритмов сегментации изображений ДЗЗ в условиях ограниченного времени обработки. Известные алгоритмы сегментации, основанные на анализе гистограмм, разделении и слиянии областей, выращивании областей не предусматривают возможности прерывания процесса обработки из-за ограничения времени вычислений. Поэтому их остановка приводит к частичной сегментации изображения, когда часть изображения полностью или частично сегментирована, а часть вообще не обрабатывалась. Такой результат не может быть использован для поиска или идентификации объектов изображений ДЗЗ с заданной точностью.

Для алгоритмов двухпорогового (2Th) и гистограммного (HTh) квантования на рисунках 2 – 9 приведены зависимости числа сегментов и среднеквадратической ошибки восстановле-

ния тестовых изображений после сегментации от времени сегментации для известных алгоритмов сегментации на основе выращивания областей (RG), блочного волнового выращивания областей изображения на основе квадросеток пикселей (узловых – NcWRG и сплошных – NnWRG).

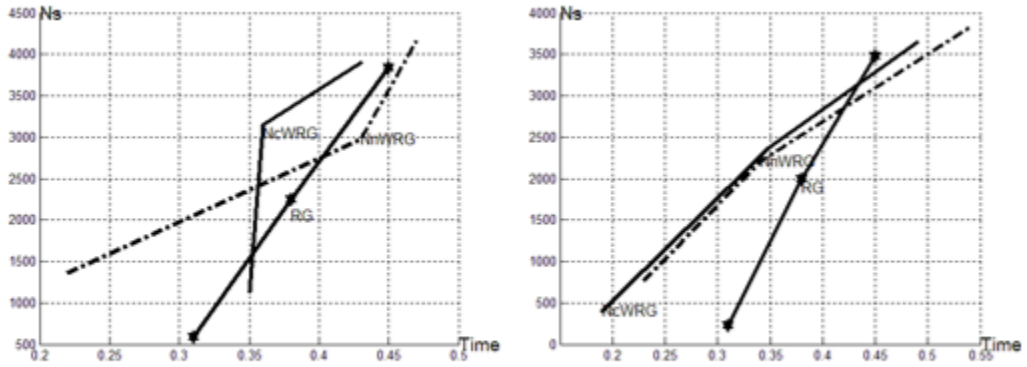


Рис. 2. Зависимости числа сегментов тестовых изображений от времени сегментации: а – при двухпороговом квантовании, б – при гистограммном квантовании изображения «Город»

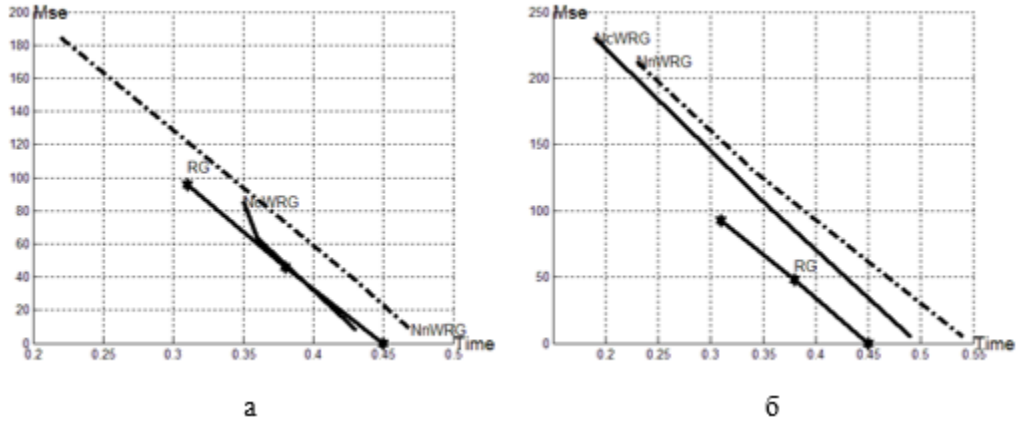


Рис. 3. Зависимости среднеквадратической ошибки тестовых изображений от времени сегментации: а – при двухпороговом квантовании, б – при гистограммном квантовании изображения «Город»

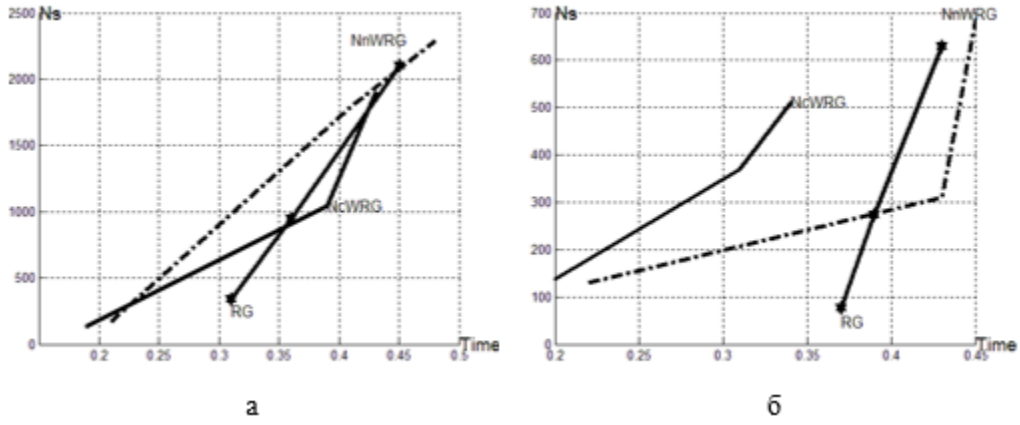
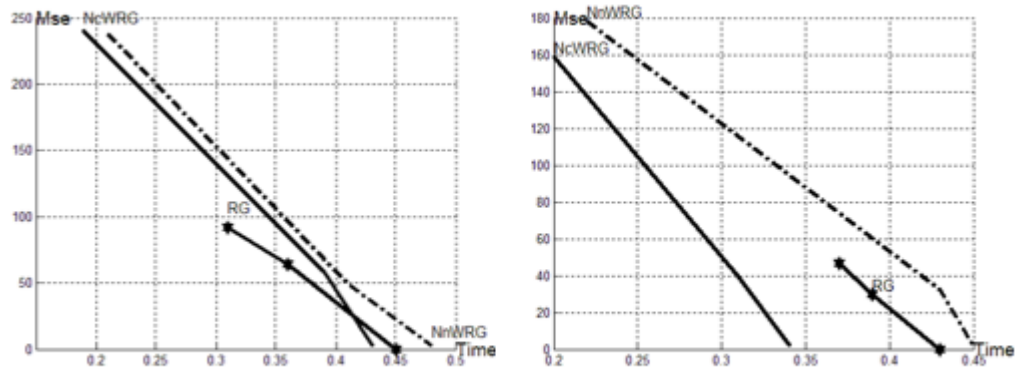
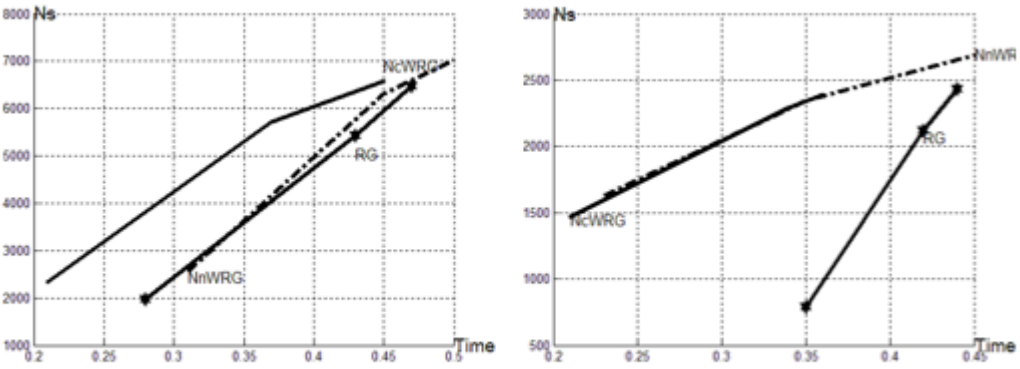


Рис. 4. Зависимости числа сегментов тестовых изображений от времени сегментации: а – при двухпороговом квантовании, б – при гистограммном квантовании изображения «Лес»



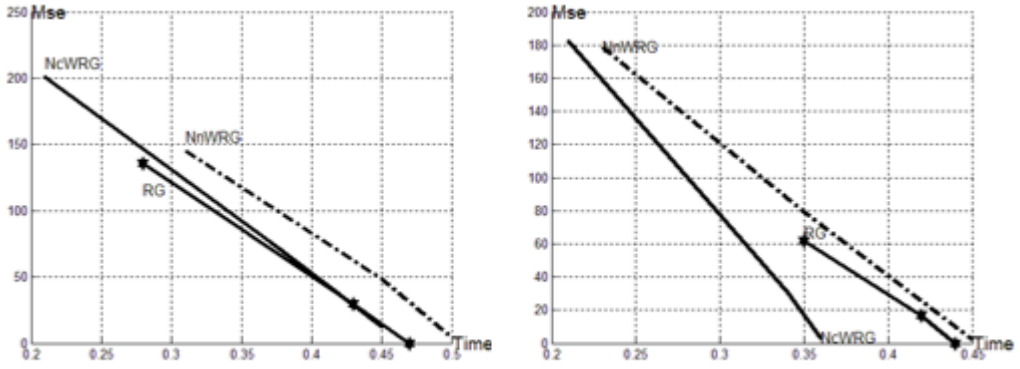
а б

Рис. 5. Зависимости среднеквадратической ошибки тестовых изображений от времени сегментации: а – при двухпороговом квантовании, б – при гистограммном квантовании изображения «Лес»



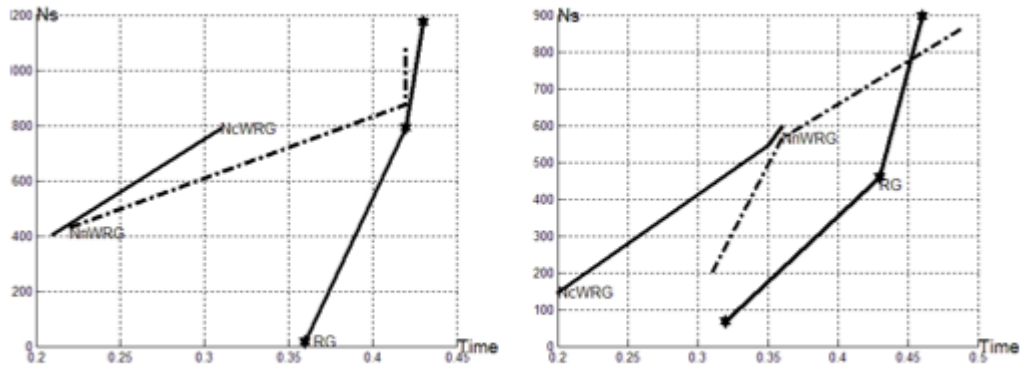
а б

Рис. 6. Зависимости числа сегментов тестовых изображений от времени сегментации: а – при двухпороговом квантовании, б – при гистограммном квантовании изображения «Степь»



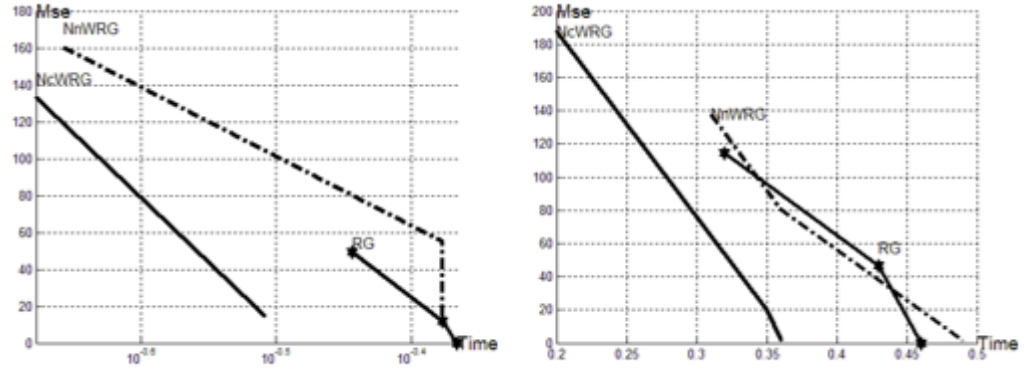
а б

Рис. 7. Зависимости среднеквадратической ошибки тестовых изображений от времени сегментации: а – при двухпороговом квантовании, б – при гистограммном квантовании изображения «Степь»



а б

Рис. 8. Зависимости числа сегментов тестовых изображений от времени сегментации: а – при двухпороговом квантовании, б – при гистограммном квантовании изображения «Горы»



а б

Рис. 9. Зависимости среднеквадратической ошибки тестовых изображений от времени сегментации: а – при двухпороговом квантовании, б – при гистограммном квантовании изображения «Горы»

Закючение. Таким образом, эксперименты для различных типов изображений ДЗЗ показывают, что в условиях ограниченного времени сегментации алгоритмы блочного волнового выращивания областей изображения на основе квадросеток пикселей эффективнее базового алгоритма выращивания областей в 1,8 и 2,8 раз по времени при двухпороговом и гистограммном квантовании соответственно.

Литература

[1]. Solomon C., Breckon T. // John Wiley & Sons, 2011. – P. 263–286.

[2]. Hyunki R., HaengSuk L. // International Journal of Software Engineering and Its Applications. – 2013. – Vol. 7. – P. 99–112.

[3]. Gill H.K., Kaur A.G.J. // International Conference on Medical Imaging, m-Health and Emerging Communication Systems (MedCom). – 2014. – P. 454–457.

[4]. Shan Y., Tsai K., Wu J. // 4th International Conference on Biomedical Engineering and Informatics (BMEI), 2014. – P. 47–51.

[5]. Gauch J.M. // IEEE Transactions On Image Processing. – January 1999. – Vol. 8/ – № 1. – P. 69–79.

[6]. Ma, J. [et al.] // 3rd International Congress on Image and Signal Processing (CISP2010). – 2010. – P. 1396–1400.

Альмияхи О.М., Цветков В.Ю., Конопелько В.К. // Доклады БГУИР. – 2016. – № 8 (102). С. 82–88.

УЧЕБНО-ИССЛЕДОВАТЕЛЬСКАЯ СИСТЕМА ОБРАБОТКИ БОЛЬШИХ ДАННЫХ

А.И. Демидчук
Заведующий лабораторией высокопроизводительных вычислений БГУИР

Д.Ю. Перцев
Ассистент кафедры кафедрой электронных вычислительных машин БГУИР

Д.И.Самаль
Заведующий кафедрой электронных вычислительных машин БГУИР, кандидат технических наук, доцент

Белорусский государственный университет информатики и радиоэлектроники, Республика Беларусь
E-mail: samal@bsuir.by.

Abstract. Educational and research system to merge different frameworks to common interface is introduced. The description of class diagram is provided to apply for this purpose.

На сегодняшний день существует огромное число проектов по анализу и обработке в области Big Data (TensorFlow[1], Theano[2], множество проектов на Python, C++), каждый из которых имеет свои плюсы и минусы. Например, реализации на Python не поддерживают работу с графическим процессором, но включают большое число уже готовых к использованию алгоритмов. Проекты TensorFlow[1], Theano[2] наоборот включают ограниченное число уже готовых алгоритмов, но предоставляют интерфейс для авторских разработок, кроме того, поддерживается обработка вычислений на GPU.

Целью работы, проводимой в рамках научно-исследовательской лаборатории в БГУИР, является разработка учебно-исследовательской системы, позволяющей унифицировать интерфейс доступа к алгоритмам анализа данных, предоставить возможность динамического построения цепочки вызовов обработчиков, сбора и анализа статистики исполнения.

Общий вид программно-аппаратной платформы исполнения, представлен на рисунке 1.

В качестве клиента может выступать любой пользователь с персональным компьютером либо ноутбуком. Для упрощения организации доступа предполагается разработка специального сайта либо Eclipse-плагина, позволяющего в удобной форме сформировать последовательность операций, передать задание на кластер, получить результат и вернуть результат конечному пользователю.

Учебно-исследовательская система включает в себя следующие модули:

- интерфейс пользователя;
- сервисы;
- библиотека алгоритмов.

Клиент через сайт взаимодействует с модулем «Интерфейс пользователя». Его основной задачей является получение задания от пользователя и анализ полученных данных. При этом клиент вправе самостоятельно сформировать последовательность действий из допустимого набора либо указать задачу, которую требуется решить.

В случае, если пользователь указывает последовательность действий, задачей модуля «интерфейс пользователя» является последовательная передача управления сервисам для вы-

зова необходимых функций и формирование результата. Результат работы передается клиенту.

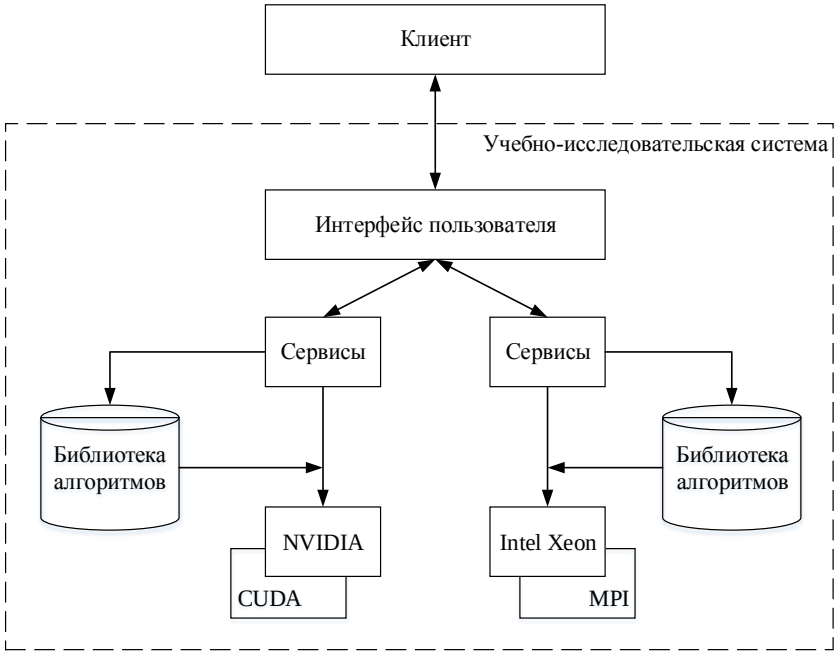


Рис. 1. Структурная схема программно-аппаратной платформы исполнения

Если пользователь указал решаемую задачу, но не сформулировал алгоритм решения, задачей модуля является формирование последовательности действий, после чего схема работы повторяет предыдущий вариант.

Обязательными требованиями к модулю «Интерфейс пользователя» являются:

– унификация интерфейса пользователя, т.к. система включает в себя множество различных компонентов с алгоритмами;

– должна быть возможность динамического обновления поддерживаемого набора функций и информирование об этом клиента. Наиболее оптимальным подходом в этом случае является применение клиент-серверной архитектуры на основе Web-технологий;

– предоставление доступа к уже готовым решениям задач (например, кластерный анализ данных), а также к отдельным алгоритмам, чтобы пользователь самостоятельно мог сформировать последовательность действий.

К ограничениям для данной модели можно отнести:

– минимальный набор поддерживаемых функций на начальном этапе. Это позволит протестировать и отладить систему, разработать оптимальную схему доступа;

– версия используемого программного обеспечения, на которой гарантируется работоспособность системы;

– информирование пользователя о поддерживаемом наборе функций предполагается осуществлять через XML-структуру, передаваемую по TCP протоколу.

В основе учебно-исследовательской системы располагается вычислительный кластер БГУИР, имеющий следующую конфигурацию:

1 вычислительный блок (7 модулей), каждый из которых включает следующие элементы:
   – блейд-стойка: GPU SuperBlade SBI-7127RG;
   – центральный процессор: Intel Xeon E5-2650 (2 шт.);
   – ОЗУ: 32 Gb RAM стандарта DDR3;

– графический процессор: NVIDIA Tesla M2075 6 Gb RAM (2 шт.);
– шина InfiniBand 4x QDR (40Gbps);
– сетевой контроллер: Gigabit Ethernet (2 шт.);

2 управляющий блок, включающий следующие элементы:

– блэйд-стойка: Processor Blade SBI-7126T-S6;
– центральный процессор: Intel Xeon E5606 (2 шт.);
– ОЗУ: 24 Gb RAM стандарта DDR3;
– жесткий диск: SSD 80Gb (2 шт.) и HDD 300Gb (4 шт.);
– шина InfiniBand 4x QDR (40Gbps);
– сетевой контроллер: 2x Gigabit Ethernet.

Основным преимуществом использования вычислительного кластера в качестве устройства исполнения являются достаточные вычислительные ресурсы для того, чтобы справиться с серьезной нагрузкой, в т.ч. длительной.

Задачей модуля «Сервисы» является контроль и управление процессом исполнения. Данные и алгоритм для исполнения передается модулем «Интерфейс пользователя», после этого формируется запрос в базу алгоритмов и передается управление, полученный результат обработки возвращается на уровень «Интерфейса пользователя».

Предлагаемая архитектура является универсальной, легко расширяемой и адаптируемой под заданные условия.

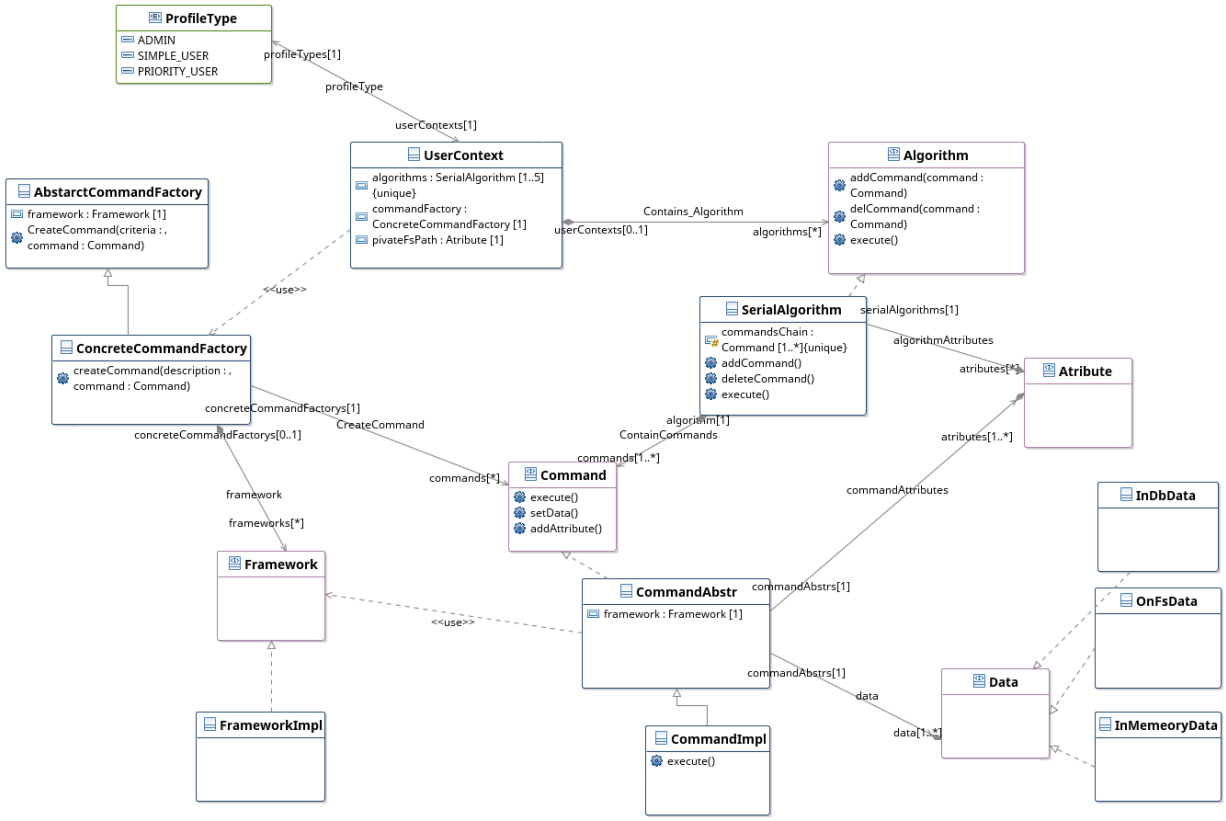


Рис. 2. Диаграмма классов для реализации учебно-исследовательской системы

На основе структурной схемы разработана диаграмма классов (рисунок 2), включающая 4 основных функциональных модуля:

–контекст пользователя;

−промежуточное звено, связывающее библиотеку алгоритмов с контекстом пользователя;

−модуль подключаемой библиотеки алгоритмов;

−модуль загрузки/сохранения данных.

Контекст пользователя включает базовый класс *UserContext*, через который осуществляется взаимодействие пользователя с библиотекой, и *ProfileType*, определяющий права пользователя.

При инициализации системы формируется все множество поддерживаемых алгоритмов через класс *ConcreteCommandFactory*. Задачей данного класса является:

– формирование статического списка команд *Command*, поддерживаемых разрабатываемой системой;

– формирование связи с необходимыми библиотеками *Framework* через *ConcreteCommandFactory*.

Каждая подключаемая библиотека представляет собой отдельный класс *Framework*, позволяющий подключить библиотеку, провести базовые настройки без участия пользователя, предоставить список поддерживаемых алгоритмов. Каждая подключаемая библиотека наследуется от *FrameworkImpl* для стандартизации доступа.

При запросе на обработку данных через *UserContext* выполняются следующие действия:

– обращение к *ConcreteCommandFactory* для получения ссылок на объекты Command, необходимые для исполнения в соответствии с указаниями пользователя;

– формирование объекта *Algorithm*, представляющего собой последовательность команд *Command* для исполнения.

По мере формирования списка команд и указания необходимых параметров через класс *Attribute* начинается обработка данных и формирование результата.

Модуль загрузки и сохранения данных *Data* предоставляет унифицированный интерфейс доступа для доступа к различным репозиториям с данными. В качестве источников данных предполагаются:

– база данных (класс *InDbData*);

– файл на жестком диске (класс *OnFsData*);

– ОЗУ (класс *InMemoryData*).

Представленная учебно-исследовательская система является универсальной, легко расширяемой и адаптируемой под необходимые условия работы.

Литература

[1]. TensorFlow [Electronic resource]. – Mode of access: https://www.tensorflow.org/. – Date of access: 01.03.2017.

[2]. Theano [Electronic resource]. – Mode of access: http://deeplearning.net/software/theano/. – Date of access: 01.03.2017.

МОДУЛЬ ПОЛУЧЕНИЯ ДАННЫХ ИЗ ВНЕШНИХ ОТКРЫТЫХ ИСТОЧНИКОВ

М.В. Стержанов
Ассистент кафедры информатики БГУИР

Д.Н. Рожков
Студент кафедры информатики БГУИР

В.Ю. Пресняцкий
Студент кафедры информатики БГУИР

П.Е. Дорошкевич
Студент кафедры информатики БГУИР

А.И. Свито
Студент кафедры информатики БГУИР

Белорусский государственный университет информатики и радиоэлектроники, Республика Беларусь
E-mail: sterjanov@bsuir.by, rdimon2912@gmail.com, presniatski@gmail.com, dpavluha@gmail.com, alexandervirk@gmail.com

Abstract. Web-crawlers (also known as robots or scrapers) enable the process by following the hyperlinks in web pages to automatically download a fractional snapshot of the web site. This paper describes developed web crawler named MMScraper aimed to process informational web resources for further getting statistical properties and performing data analytics.

В настоящее время в связи с бурным развитием сети Интернет наблюдается обилие электронной неструктурированной информации, представленной текстами на естественных языках. Всё более востребованной становится задача автоматической обработки таких текстов с целью извлечения структурированных данных, которые затем используются при решении различного рода проблем: извлечения фактических данных, поиска информации и т.п. Нами решается задача обработки контента информационно-новостных ресурсов с целью анализа лексико-терминологической информации.

Для сбора требуемых данных требуются специализированные инструменты - поисковые роботы, также называемые «веб-пауками» (web-spider), краулерами (web crawler) или скребками (web scraper). Поисковый робот — программный комплекс, осуществляющий навигацию по веб-ресурсам и сбор информации для базы данных приложения-агента [1, 2].

Нами планируется значительная работа по обследованию ряда информационных сайтов, чтобы собрать выборку данных требуемого размера. Анализ имеющихся в свободном доступе решений показал, что открытые реализации зарубежных веб-краулеров слабо приспособлены к решаемой нами задаче, так как требуют весьма трудоемкой настройки, а после нее показывают низкую производительность и существенно нагружают информационный источник. В связи с этим было принято решение разработать собственное решение.

Опишем основные требования, в соответствии с которыми был разработан краулер,

названный MMScraper.

– Получать в качестве исходных данных список доменных имен сайтов, предназначенных для сканирования. Предполагается, что имеется некоторое множество заранее определенных для исследования сайтов.

– Обходить каждый сайт, начиная с главной (индексной) страницы, перемещаясь по внутренним гиперссылкам в заданном порядке обхода «вначале вширь».

– Полученные результаты сохранять в базу данных. Интерес представляют следующие атрибуты: адрес страницы, автор публикации, дата публикации, содержимое публикации.

– Позволять получать данные с сайта через программный интерфейс API.

– Иметь расширяемую архитектуру для последующего развития функциональности.

– Добавление нового сайта должно быть простым и не требовать привлечения квалифицированного программиста.

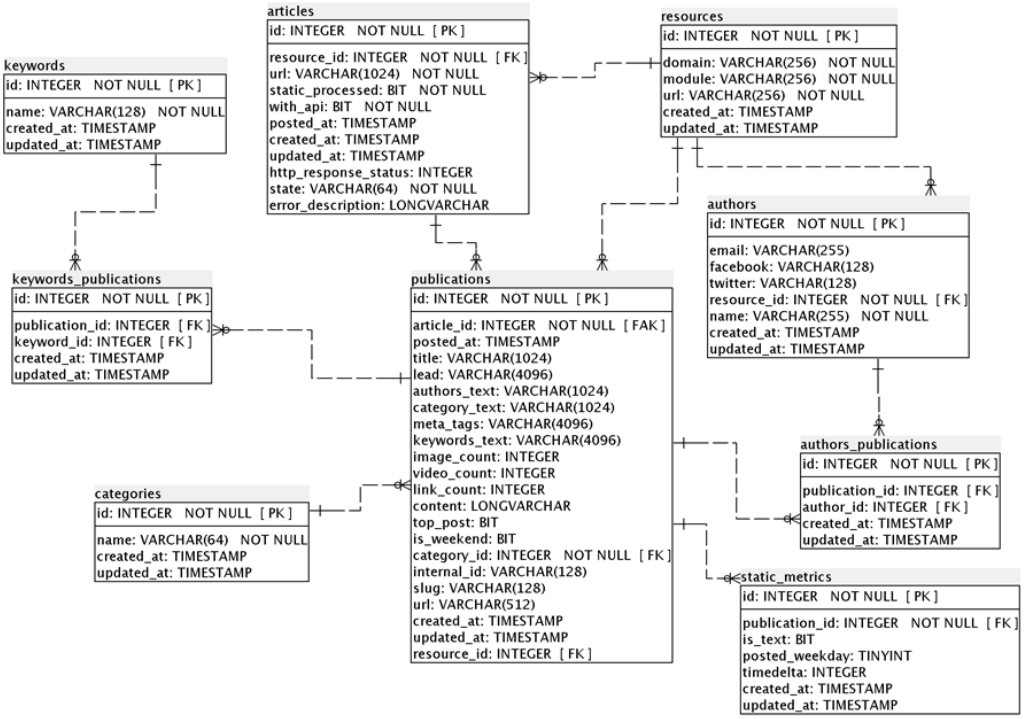


Рис. 1. Модель базы данных

Работу разработанного краулера можно описать следующим образом: сканирование сайта начинается с начальной страницы и затем робот использует ссылки, размещенные на ней, для перехода на другие страницы. Каждая страница сайта анализируется на наличие требуемой информации, которая копируется в соответствующее хранилище в случае обнаружения. Процесс повторяется до тех пор, пока не будет проанализировано требуемое число страниц либо пока не будет достигнута некая цель. Модуль получения данных разработан на языке программирования Ruby и состоит из трех основных частей: блок сканирования и обработки данных, блок управления краулером (команды вводятся через консоль) и база данных. Собираемая роботом информация состоит из ссылочной структуры обрабатываемого ресурса и веб-страниц. В качестве основы для базы данных была выбрана бесплатная СУБД MySQL. Для упрощения взаимодействия с БД нами используется библиотека Sequel, позволяющая представлять данные в виде объектов.

Рассмотрим схему базы данных, содержащую полученную информацию.

Таблица *resources* описывает веб-сайты, которые подлежат краулингу. Атрибут *module* сообщает какой шаблон отвечает за разбор полученной страницы и выделения полей, необходимых для сохранения в БД.

Таблица *articles* содержит ссылки на страницы сайта, которые подлежат скачиванию. Таблица *publications* представляет информацию, полученную путем разбора целевых страниц сайта. Как видно их схемы данных, мы храним заголовок, аннотацию и текстовое содержание документа. Каждая публикация принадлежит категории (таблица categories), имеет ключевые слова (таблица keywords), и написана одним или несколькими авторами (таблица authors). Помимо этого, мы подсчитываем число изображений, видео и ссылок, содержащихся на странице.

В работе приводится описание основных требований, общей архитектуры и конфигурации краулера MMScraper, предназначенного для решения достаточно узкой, но важной задачи, а именно – сбора информации о новостных и информационно-аналитических публикациях.

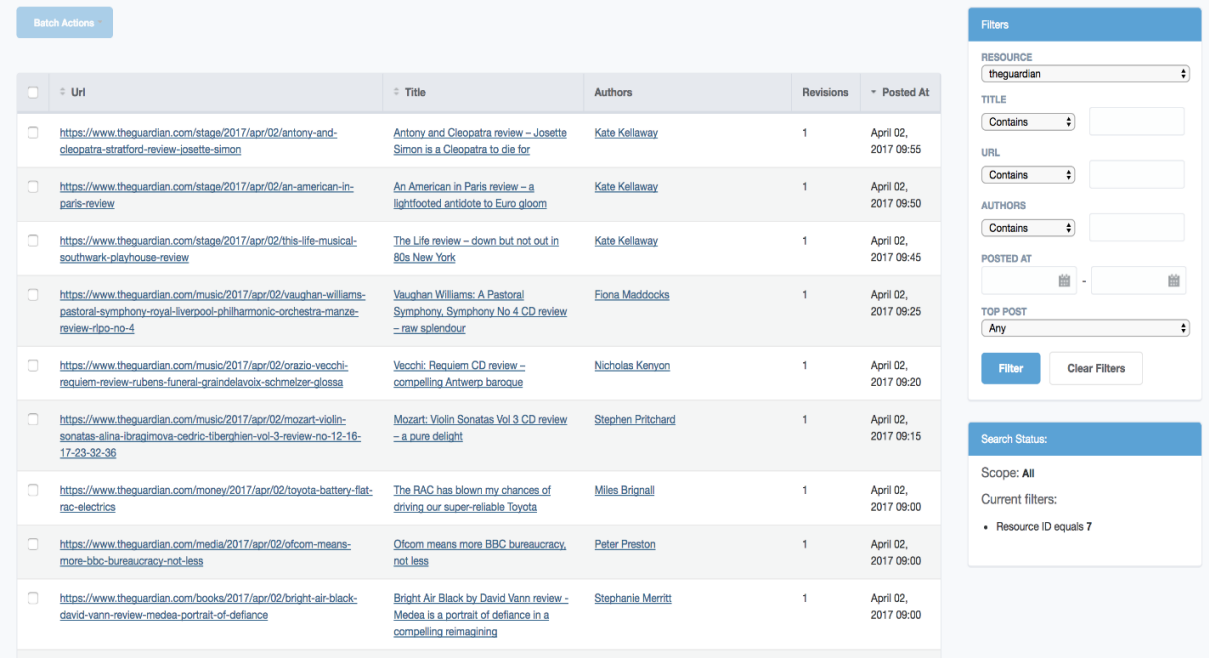


Рис. 2. Графический интерфейс пользователя

В практическом плане разработанный MMScraper позволит собрать экспериментальную базу для исследования задач интеллектуальной обработки текста.

Нам видится важным продолжить исследования результатов краулинга и реализации дополнительных возможностей MMScraper, улучшающих результаты работы на очень больших сайтах. Реализация таких возможностей предусмотрена расширяемой архитектурой краулера.

Литература

[1]. A.H.F. Laender, B. A. Ribeiro-Neto, Juliana S.Teixeria. A brief survey of web data extraction tools // ACM SIGMOD Record 31(2), pp 84-93. 2002

[2]. Baeza-Yates R., Castillo C. Crawling the Infinite Web: Five Levels are Enough // Lecture Notes in Computer Science. Algorithms and Models for the Web-Graph, Third International Workshop. 2004. Vol. 3243. P. 156–167.

ПРИМЕНЕНИЕ ТЕХНОЛОГИЙ BIG DATA В СФЕРЕ ТРАНСПОРТА

А.А. Александров
Аспирант кафедры информатики БГУИР

И.И. Пилецкий
Научный руководитель совместной лаборатории БГУИР-IBA и АЦКТ IBM, архитектор отделения по программному обеспечению IBA IT Park, доцент кафедры информатики БГУИР, кандидат физико-математических наук

Белорусский государственный университет информатики и радиоэлектроники, Республика Беларусь
E-mail: zxbyteman@gmail.com.

Abstract. This paper describes some use cases of Big Data technologies for the sphere of transport, such as schedule planning, failure prediction of transport equipment based on diagnostic and vibration data, issues with transmitting of big amounts of data to remote datacenter using slow data transmitting channels.

Использование технологий Big Data в сфере транспорта позволяет решить различные нетривиальные задачи в сфере транспорта.

Одной из проблем использования технологий Big Data на транспорте является проблема доставки больших данных в центры обработки данных. Если для городского транспорта, где покрытие мобильными сетями 3G приближается к 100%, проблем в отправке больших объёмов данных нет (данные можно накапливать на борту транспортного средства и передавать их в другое время, например, когда транспорт находится в парке), то для транспорта, такого как грузовые автомобили, поезда, грузовые самолёты, есть определенная проблема в доставке больших объёмов данных.

В местах недостаточного уровня сигнала для качественной передачи данных между базовой станцией оператора сотовой связи и установленным оборудованием, в блоке передачи предусмотрена «буферизация» (запись) данных на внутреннюю память. В момент появления устойчивой связи с качественным сигналом «буферизованные» данные передаются на сервер в автоматическом режиме.

Поскольку зачастую для достижения поставленных задач необходима высокая частота дискретизации снимаемых параметров (например, мировые координаты, моментальный расход топлива, состояние транспортного средства, состояние агрегатов транспортного средства и др.), существует проблема разработки сложных алгоритмов сжатия и регулярной доставки этой информации или, когда появляется стабильный канал связи.

Какую информацию можно собирать? В первую очередь это точные координаты с привязкой ко времени. На основании этой информации можно производить планирование и составление точных графиков и расписаний движения транспорта. В Минске в качестве пилотного проекта реализована система управления светофором на перекрестке улиц Козлова и Платонова на базе RFID-меток. Данную систему можно усовершенствовать, используя телеметрию с трамвая для упреждающего управления светофором. Система может подстроить ра-

боту светофора таким образом, чтобы минимизировать задержку трамвая на перекрестке и переходить к следующей фазе работы светофора сразу после прохода трамвая через перекресток.

В качестве передаваемых данных для локомотивов минимально должны быть: геокоординаты нахождения локомотива и точное время; серия, номер и депо приписки локомотива, на котором установлено оборудование; скорость перемещения локомотива.

Текущая парадигма ИТ систем, эксплуатируемых на железных дорогах – это *сопровождение процесса перевозок и эксплуатации железнодорожных объектов* [1].

Данные факты являются серьезной проблемой для создания среды интеллектуализации железных дорог, которая может быть выражена как: анализ спроса погрузки и доставки груза, анализ пропускной железнодорожной сети, определения маршрутов, составления расписания и регулирования графиков движения поездов, контроля состояния подвижного состава (локомотивов, вагонов) и диагностирование поломки подвижного состава, увеличения пропускной способности существующих линий (за счет точных знаний координат ПС, сокращения дистанции между ПС, увеличения скорости движения), решения задач транспортной логистики (в том числе и предоставление полной информации о грузе и его транспортировке грузовладельцам) и др.

Новая парадигма ИТ систем на железных дорогах – *интеллектуальные системы прогнозного типа, анализа, выбора и принятия оптимальных и упреждающих решений,* т.е. наряду с существующими технологиями, требуются интеллектуальные системы, предоставляющие полный цикл логистических услуг, обеспечивающие анализ и выбор оптимальных вариантов и решений опережающих события, безопасность транспортировки грузов и эксплуатации транспортных средств.

Основными факторами интеллектуализации железных дорог являются:

– технологическая оснащенность: идентификация подвижного состава, груза, точное знание геокоординат подвижного состава и груза на геополигоне дороги в режиме real-time;

– беспроводные системы мониторинга, позволяющие отслеживать актуальные координаты подвижного состава и грузов, находящихся в пути, и маршруты их следования;

– информация о техническом состоянии подвижного состава и ситуации сети железных дорог (на перегонах, на станциях) в режиме, близкому к real-time;

– интеграция взаимодействия компонент новых и уже существующих, обработка всех железнодорожных документов в центральных серверах, минимизируя обработку данных на линейном уровне;

– управленческие обоснованные решения на основе анализа больших и сверх больших объемов данных, поступающих со всей дороги – принятие оперативных решений в режиме real-time, а далее тактических и стратегических решений.

Так, сбор телеметрической информации текущего состояния двигателя позволяет принимать как оперативные, так и стратегические решения. Например, применительно к автобусу, такой информацией являются данные о моментальном расходе топлива, оборотах двигателя, текущая передача, нагрузка на ось и т.д. Обычно анализируют лишь данные ошибок двигателя, игнорируя походную информацию. Однако, при фиксации этих данных с высокой частотой дискретизации, можно провести анализ этих данных и выявить сложные режимы работы двигателя, оптимизировать конфигурацию двигателя для минимизации расхода топлива и обеспечения оптимального режима работы двигателя.

Также эти данные могут помочь для расчета текущего пассажиропотока на основании данных о текущей нагрузке на ось и выявлять высоконагруженные маршруты с дальнейшим принятием решений о корректировке графика движения и увеличения числа единиц подвижного состава на данном маршруте.

Малоизученной сферой применения Big Data и предиктивной аналитики на транспорте является предсказание неисправностей агрегатов и механизмов на основании данных объективного контроля и вибромониторинга. Собрав большую базу штампов работы агрегатов,

можно с высокой долей вероятности фиксировать аномалии работы двигателя и предугадывать выход из строя того или иного агрегата. Учитывая большие вычислительные мощности ЦОД и развитие технологий обработки такого рода данных можно получить информацию об отказе агрегата с точностью до движущейся детали. Объёмы сырых данных исчисляются гигабайтами в час, поэтому есть необходимость в разработке алгоритмов, позволяющих оптимизировать передаваемый поток сырых данных с датчиков таким образом, чтобы не потерять в достоверности передаваемых сэмплов и не потерять важную информацию при сжатии. Так, например, в 2015 только локомотивов на БелЖД было 920 единиц, при активном использовании только 70% прописного парка, передаче только минимальных данных о локомотиве (100 байт), каждые 10 секунд обьем передаваемых данных будет не менее 556.041.600 байт в сутки. Здесь не учтен большой объём данных о работе агрегатов локомотива и расходе топлива, данные о составе поезда, данные о бригаде, а также обратно передаваемые данные на локомотив.

Новые технологии идентификации объектов, сбора и передачи информации позволяют уточнить традиционно решаемые задачи:

– *Уровня операционного управления,* которые требуют принятия правильных решений в режиме близкому к real-time.

– *Уровня текущего или тактического управления,* которые требует поиска оптимальных вариантов решения часто возникающих типовых задач. Данные используются те же что были собраны ранее, на уровне операционного управления, но размещенные в базах данных и хранилищах железнодорожных систем.

– *Уровня стратегического управления,* которые требует выработки планов деятельности на основе качественных прогнозов и различных видов анализа.

Проанализировав вышесказанное, можно выделить несколько направлений для исследований:

– Использование Big Data для решения проблемы составления расписаний и графиков движений транспорта (городского, дальнемагистрального, железнодорожного). Сюда же входит вопрос оптимизации расходов при эксплуатации транспорта с учетом рельефа местности (например, для грузовых составов поездов вес которых, может превышать 4000 тонн, весьма критичным вычисление текущей скорости состава с точностью до 0,1км/ч, на поворотах, подьемах/спусках, линейных участках, перед светофором, что может позволить сэкономить топливо/электроэнергию на разгон и торможение до 10-15%).

– Исследование применения Big Data для предиктивной диагностики и мониторинга состояния транспорта и подвижного состава, разработка программно-аппаратных комплексов для предсказания выхода из строя тех или иных агрегатов, сбор и анализ больших массивов данных с физических датчиков вибрации, температуры и т.д., что позволит отказаться от плановых ремонтов ТО1, ТО2 и др., а перейти непосредственно к требуемому обслуживанию подвижного состава.

– Исследование методов передачи больших массивов данных в условиях узкого канала, методики сжатия с учетом специфики данных, минимизацию задержек от отправки данных до получения обратного ответа с рекомендациями о параметрах движения, техобслуживания и т.д.

В рамках программы исследования загруженности городских автобусов и перспективы замены их на гибридные электробусы было разработано устройство, устанавливаемое в моторный отсек автобуса и подключаемое к внутренней CAN-шине контроллера двигателя Mercedes ADM3. Устройство было предназначено для непрерывного сбора информации с высокой частотой о работе двигателя с привязкой к текущему времени. Информация записывалась на SD-карту, раз в неделю инженеры производили считывание информации и передачу её для дальнейшего анализа в лабораторию. Устройство «каталось» на протяжении нескольких недель на автобусном маршруте №18 г. Минска. Результатом анализа полученного массива данных явилось решение о целесообразности перехода на гибридные двигательные установки. Решение основывалось на анализе режимов работы двигателя в городских условиях и нагрузки

на двигатель, времени простоев на остановках, светофорах, динамики разгона-остановки и прочих факторов.

На данный момент ведется разработка прибора для сбора данных с транспортных средств, имеющий в себе канал связи на базе GSM-модема, привязку с текущим координатам посредством систем геопозиционирования GPS-ГЛОНАСС, при этом имеющим минимальные габариты и возможность работать в агрессивных условиях машинных отделений.

ЦОД лаборатории БГУИР-IBA [2], может использоваться как база в пилот проектах для хранения и обработки данных с датчиков.

Наиболее перспективным направлением является применение сервисов платформы IBM Bluemix и когнитивного суперкомпьютера IBM Watson [3, 4], возможности которого позволяют совместить обработки больших объёмов данных, получаемых с различных датчиков. IBM Bluemix предоставляет специальные модули для получения информации с IoT-устройств, к которым можно отнести разрабатываемый автономный прибор сбора данных.

Когнитивные системы, приложения и сервисы, аналитика (Watson) + IoT (Internet of Things) – Интернет вещей (например, автомобили, локомотивы, полигон дороги), позволяют с минимальными усилиями создать сеть взаимодействия M2M (Machine-to-Machine), что в будущем обеспечит переход к безлюдному производству.

Литература

[1]. Пилецкий И. И. «Один из методов построения и модернизации корпоративных приложений» Материалы конференции - “Software Engineering Conference (Russia) SEC(R) 2007”, Moscow, November 1-2 , 2007.

[2]. И.И. Пилецкий и др. Виртуальная ИТ среда БГУИР для исследования Big Data и VCL, с. 21-32, BIG DATA and Predictive Analytics. Использование BIG DATA для оптимизации бизнеса и информационных технологий : сборник материалов междунар. науч.-практ. конф. / редкол. : М.П. Батура [и др.]. – Минск : БГУИР, 2015. – 220 с. ISBN 978-985-543-146-7. - С. 21-32.

[3]. What is Bluemix [Электронный ресурс] / IBM developerWorks. – 2015-2017. – Режим доступа: https://www.ibm.com/developerworks/cloud/library/cl-bluemixfoundry/. – Дата доступа: 20.03.2017.

[4]. IBM RCIS Watson Cloud Cognitive University [Электронный ресурс] / IBM Developer Works. – 2016-2017. – Режим доступа: https://www.ibm.com/developerworks/community/groups/service/html/communitystart?communityUuid=bc004137-b64a-4378-ac02-2caf59c56c2a. – Дата доступа: 20.03.2017.

ТЕХНОЛОГИИ СЕМАНТИЧЕСКОЙ ОБРАБОТКИ ИНФОРМАЦИИ В УЧЕБНОМ ПРОЦЕССЕ

Л.Ю. Шилин
Декан факультета информационных технологий и управления БГУИР, доктор технических наук, профессор

А.А. Навроцкий
Заведующий кафедрой информационных технологий автоматизированных систем БГУИР, кандидат физико-математических наук, доцент

Л.С. Стригалев
Старший преподаватель кафедра информационных технологий автоматизированных систем БГУИР

Белорусский государственный университет информатики и радиоэлектроники, Республика Беларусь
E-mail: dekfitu@bsuir.by, navrotsky@bsuir.by, orion@bsuir.by

Absract. Обсуждаются методологические вопросы применения технологий семантической обработки информации в учебном процессе.

Сложность, высокая динамичность развития и стоимость современных технологий порождают многочисленную проблематику, предъявляя повышенные требования к образовательным технологиям, особенно в IT-сфере, которая имеет не только всевозрастающую долю ВВП в мировой экономике, но и в силу своего конвергентного характера порождает дополнительные специфические проблемы, в том числе и методологические проблемы подготовки IT-специалистов.

Важным и одновременно сложным «срезом» образовательной технологии становится адекватный мониторинг качества работы студента, необходимый для его оптимальной проводки в образовательном пространстве. В современных условиях академических свобод и использования традиционных технологий, включая «облачные» [1−4], для эффективной реализации образовательных траекторий явно недостаточно. К тому же необходим и современный конструктивный методологический аппарат, поскольку традиционная образовательная парадигма не только не отображает системную сложность образовательных технологий и объектов, но и не находит адекватного места для систем и средств компьютерного интеллекта, которые в ряде областей достигли уровня «интеллекта» среднего человека.

Объекты образовательной среды (обучаемый, обучающий и любая обучающая организация) могут быть описаны четверкой: система, структура, цель, технология, с тремя взаимосвязанными уровнями целеполагания: генетический, неосознанный и осознанный [2, 3].

Обучение человека осуществляется на неосознанном и осознанном уровне целеполагания как информационно-энергетическая составляющая метаболизма, которая имеет три технологических уровня: синтаксический, семантический и прагматический. Синтаксический уровень (восприятие, преобразование, передача, хранение информации) в обучении носит обеспечивающий характер; в технических системах это по большей части хорошо разработанный инфраструктурный уровень.

Важнейшим уровнем в обучении является неосознанный уровень (условный и каузальный рефлексы; ментальность, привычка). На этом уровне в процессе обучения формируется общая культура и профессиональные компетенции. Осознанная целенаправленность, на

уровне которой формируется образовательная траектория, играет активную роль в формировании личности и совместно с внешней средой воздействует на структуру неосознанной целенаправленности. Пространство свободы человека, который обучается практически в течение всей жизни (но фундамент интеллекта закладывается в когнитивном возрасте, до 17-18 лет) определяется генетическим и неосознанным уровнями, а также состоянием внешней среды.

Так как процесс обучения реализуется на семантическом технологическом уровне, то учебный материал и результаты обучения носят структурированный характер и, следовательно, имеют конечную колмогоровскую сложность, что позволяет автоматизировать оценку качества процесса обучения. Колмогоровская сложность характеризует индивидуальные свойства объекта, чем она меньше, тем выше структурированность объекта. Структурированный объект с вероятностью близкой к единице порождает эргодические (типичные) реализации. Определенным аналогом колмогоровской сложности (но, по множеству реализаций) является шенноновская избыточность.

Неструктурированный объект, имеющий нулевую избыточность (максимальную колмогоровскую сложность), не имеет семантики. Наличие избыточности свидетельствует о существовании закономерности (структуры), порожденной либо интеллектуальной системой, либо объектом с известными или непознанными свойствами. Очевидно, что система (естественная или искусственная) способная обнаруживать, анализировать и «работать» с семантикой является интеллектуальной, в особенности, если она способна еще и принимать решения. Подобного рода системы становятся все более востребованными, что напрямую связано с понятием «интернет вещей» (Internet of Things, IoT). По мнению компании Cisco Интернет вещей это момент времени, когда количество материальных объектов, подключенных к Интернету, превысило количество людей, пользующихся Интернет; что произошло в промежутке между 2008 и 2009 годами.

Особую роль в деле семантической обработки информации играют технологии Data Mining. Понятие Data Mining, появилось в 1978 году; популярность приобрело примерно в первой половине 1990-х годов. Особенно актуальны технологии Data Mining в настоящее время, что обусловлено всевозрастающими объемами неструктурированной информации, которую порождает не только человек, традиционные технологические и научно-исследовательские средства и системы, но и многочисленные мобильные устройства, количество которых значительно превышает население Земли. В результате возник технологический разрыв между возможностями и потребностями в обработке семантических свойств информации, который потребовал развития адекватных средств искусственного интеллекта, реализующих семантическую обработку. Возможности средств искусственного интеллекта на данный момент в ряде случаев соответствуют уровню среднестатистического человека.

О роли и назначении Data Mining в обработке информации достаточно красноречиво говорят варианты перевода этого термина: добыча данных, "раскопка" данных, извлечение знаний, извлечение "зерен знаний" из гор данных, "промывание" данных, информационная проходка данных, интеллектуальный анализ данных и т. д. Сказанное, по существу, позиционирует и многочисленные сферы применения Data Mining.

Технологии Data Mining, которые по существу входят в состав методов анализа больших данных (Big Data [5]), развивалась на базе прикладной статистики (это стартовая позиция данной технологии), теории искусственного интеллекта и машинного обучения, теория баз данных и других областей. Следует заметить, что в ряде источников технологии методов Big Data и Data Mining пересекаются, например, машинное обучение, распознавание образов и др.

Технологии Data Mining можно классифицировать по многочисленным признакам. По типу обрабатываемых данных Data Mining подразделяются:

- Text Mining — технологии поиска и семантического анализа текста;
- Web Mining — интеллектуальный анализ данных в Internet;
- Call Mining — «добыча звонков», технология распознавания речи и ее анализ;

– Audio Mining — извлечение данных из аудиозаписей;
– Video Mining — извлечение данных из видеозаписей.

Data Mining имеет неограниченную, постоянно расширяющуюся сферу применения, однако наибольший эффект, который, как отмечается в ряде источников, может достигать 1000% характерен для коммерческих предприятий, что в конкурентной борьбе делает неотвратимым применение данной технологии. Кроме коммерческой сферы технологии Data Mining находят применение в производстве, телекоммуникации, государственном управлении, научных исследованиях, медицине, геологоразведке и т.д.

Одним из перспективных направлений Data Mining является технологии Educational Data Mining (EDM), которые ориентированы на исследования данных, используемых в образовательных целях, для анализа и принятия решений в сфере образования. EDM (первая конференция прошла в 2008 году в Монреале), в виду сложности образовательных технологий, имеют определенную специфику, которая помимо стандартных методов Data Mining, предполагает использование и других специфичных методов, например, методов психометрии.

В современных условиях применение технологий Data Mining в сфере образования является не только перспективным, но и необходимым как по экономическим, так и по стратегическим соображениям. Для достижения максимального эффекта эти технологии должны внедряться как в основные, так и обеспечивающие процессы учреждения образования, а также в процессы мониторинга системы качества этого учреждения. Главным направлением применения технологии EDM является оптимальная проводка студента в образовательном пространстве с максимальным использованием генетического потенциала студента.

Литература

[1]. Батура М.П. Дистанционное образование: концепция, технологии, контент, сервисы / М.П. Батура, Б.В. Никульшин, В.Ю. Цветков //Дистанционное обучение – образовательная среда XXI века: Материалы VII Междунар.научн-метод. Конференции, 1-2 декаб. 2011 г. – Минск: БГУИР, 2011 – С.7-12.

[2]. Strigalev L.S, German O.V. Methodological aspects of the IT-specialists training // Информационные технологии и системы 2011: Материалы Международной конференции, Минск, БГУИР, 2011 - С.199, 200.

[3]. Стригалев Л.С. Слабоструктурированные аспекты технологии дистанционного обучения. //Дистанционное обучение − образовательная среда XXI века: Материалы VI Междунар. научн.-метод. конференции,22-23 нояб.2007 г. Минск : БГУИР. 2007. С.230-232.

[4]. Никульшин Б.В. О создании частного образовательного облака / Б.В. Никульшин, В.Е., Проволоцкий, Е.М. Димидюк, Л.С. Стригалев /Информационные технологии и системы 2013 (ИТС 2013): Междунар.научн.-метод. конференции, 23 октября 2013. Минск: БГУИР, 2013. С. 304-305.

[5]. Шилин Л.Ю. Технология больших данных как стратегическое направление / Л.Ю. Шилин, Навроцкий А.А., Герман О. В., Л.С. Стригалев // BIG DATA and Predictive Analytics. Минск: БГУИР, 2016. С. 271-273.

МОДЕЛИРОВАНИЕ ПРОЦЕССА ОЧИСТКИ ГАЗОПЫЛЕВЫХ ПОТОКОВ В ВОЛОКНОВЫХ ФИЛЬТРАХ

М.В. Тумилович
Начальник управления подготовки научных кадров высшей квалификации БГУИР, доктор технических наук, доцент

Л.П. Пилиневич
Профессор кафедры инженерной психологии и эргономики БГУИР, доктор технических наук, профессор

Белорусский государственный университет информатики и радиоэлектроники, Республика Беларусь
E-mail: tumilovich@bsuir.by

Abstract. Modeling of sedimentation of high-disperse particles from gas streams in porous fiber materials is carried out. It is shown, under usual conditions efficiency of a kolmatation of high-disperse particles with a diameter of 0,5 microns and below, generally depends on the speed of a stream and the size of fibers of the filter – the speed more than diameter of fibers is higher, the efficiency of a kolmatation is lower. Increase in density of packing of the filtering material has not so significant effect on extent of catching of particles, but leads to growth of resistance and according to decline in production of the filter.

При производстве и механической обработке большинства материалов и изделий выделяется значительное количество пылеобразных веществ. Одним из способов снижения концентрации вредных пылеобразных веществ до значений, не превышающих предельно-допустимые нормы, является очистка воздуха с применением традиционных методов и очистного оборудования. Однако нестабильность рабочих характеристик используемого оборудования при изменении концентрации, физико-химического и дисперсного состава фильтруемых частиц, требует совершенствования конструкций и оптимизации эксплуатационных параметров фильтрующих аппаратов. Одним из наиболее эффективных методов, позволяющим провести оптимизацию эксплуатационных параметров фильтрующих аппаратов является компьютерное моделирование. Для моделирования процесса очистки газопылевых потоков с помощью современных компьютерных средств, позволяющих осуществлять выбор оптимальных режимных параметров, необходимо вначале выполнить формализацию процесса.

Целью данной работы является моделирование процесса очистки газопылевых потоков в пористых волокновых материалах (ПВМ).

Моделирование процесса очистки высокодисперсных частиц в ПВМ целесообразно начать с рассмотрения осаждения на одиночном модельном цилиндре, являющимся элементом структуры ПВМ, что позволяет перейти к исследованию процессов осаждения в реальном фильтре. Следует отметить, что волокновые фильтры имеют весьма сложную пространственную структуру, что затрудняет ее моделирование. Поэтому стандартные уравнения фильтрации не всегда являются адекватными. Это обстоятельство требует, как более глубокого изучения возможности применения уже разработанных моделей применительно к реальным фильтрам и условиям фильтрации, так и разработки новых более универсальных моделей, позволяющих с максимальной точностью предсказать эффективность работы фильтра.

На практике качество фильтра оценивается степенью выноса высокодисперсных частиц

из фильтра, равной отношению массы частицы, пропущенных фильтром к массе частиц, поступивших на него с газом, или обратной величиной − степенью задержки частиц, которая характеризует эффективность фильтра.

Примем, что однородный волоконный фильтр состоит из монодисперсных волокон, расположенных параллельно друг другу и перпендикулярно потоку, которые образуют регулярную упаковку. Будем считать, что позади волокон происходит полное перемешивание и выравнивание концентрации в поперечном сечении, а частицы не оказывают взаимного влияния друг на друга. Кроме того, в связи с тем, что ПВМ, как правило, характеризуются малой плотностью упаковки, т.е. большими расстояниями между волокнами, значительно превышающими размеры частиц, ситовым эффектом пренебрегаем.

Определим эффективность осаждения частиц в слое волоконного фильтра с учетом его параметров.

Объем волокон в фильтре площадью поперечного сечения S и толщиной d_h с учетом плотности упаковки α равен $\alpha S d_h$. В то же время объем волокон общей длины L равен $\pi d_f^2 L/4$. Отсюда длина волокна $L = 4\alpha S d_h / \pi d_f^2$. Средняя скорость частиц в фильтре с плотностью упаковки α равна $U(1-\alpha)$. Площадь сечения фильтра, занятая волокном, равна $d_f L$. Поток частиц, приходящийся на это сечение: $c d_f L U(1-\alpha)$, где c − концентрация частиц.

Доля частиц, осевших на волокнах под действием всех механизмов захвата, равна $\xi_\Sigma' d_f L U_0 c/(1-\alpha)$, где ξ_Σ' − суммарный коэффициент захвата волокном единичной длины. Та же величина убыли частиц при прохождении потока через фильтр может быть выражена в виде $-SU_0 dc$. Приравнивая эти два выражения и подставив значение L, получим:

$$-dc = \frac{4\xi_\Sigma' d_f S U_0 \alpha c dh}{U_0 \pi (1-\alpha) d_f^2 S}, \quad \text{или} \quad \frac{dc}{c} = \frac{4\alpha\xi_\Sigma'}{\pi d_f (1-\alpha)} dh \tag{1}$$

Проинтегрировав (1) в пределах высоты слоя h от 0 до H при концентрации частиц соответственно $c_{вх}$ и $c_{вых}$, получим:

$$\ln \frac{c_{вых}}{c_{вх}} = -\frac{4\alpha H \xi_\Sigma'}{\pi d_f (1-\alpha)} = \ln K \tag{2}$$

или

$$K = \frac{c_{вых}}{c_{вх}} = e^{-f\xi_\Sigma'} = \exp(-f\xi_\Sigma'),$$

где K − коэффициент проскока, f − параметр фильтрации, связанный только со структурными и геометрическими параметрами фильтра:

$$f = \frac{4\alpha H}{\pi d_f (1-\alpha)}. \tag{3}$$

Эффективность однородного фильтра толщиной H определяется в виде:

$$\xi_{\Phi} = 1 - K = 1 - e^{-f\xi'_{\Sigma}} = 1 - \exp(-f\xi'_{\Sigma}) \tag{4}$$

или эффективность осаждения по фракциям

$$\xi_{\Phi}(x) = \frac{c_{вх}(x) - c_{вых}(x)}{c_{вх}(x)} \tag{5}$$

Формула (2) содержит очевидную связь между параметрами структуры и эффективностью фильтрации, и может применяться для расчетов, если известны величины ξ'_{Σ}, α и d_f.

Плотность упаковки можно определить как отношение плотности фильтра к плотности материала волокон: $\alpha = \rho_{\Phi} / \rho_f$ – с помощью метода РЭМ (растровой электронной микроскопии).

Параметр фильтрации может быть также найден и через пористость фильтра $\Pi = 1 - \alpha$, т.е.

$$f = \frac{4}{\pi} \cdot \frac{(1-\Pi)}{\Pi} \cdot \frac{H}{d_f} \tag{6}$$

В этом случае пористость определяется одним из известных методов [1].

Задача определения структурных параметров, входящих в (3), решается более просто, если структура материала характеризуется поверхностью волокон. Это позволяет в случае реального фильтра частично избежать учета полидисперсности волокон.

Если в (3) правую часть умножить и разделить на S, получим:

$$f = \frac{4\alpha HS}{d_f} \cdot \frac{1}{S\pi(1-\alpha)} \tag{7}$$

где член $4\alpha HS / d_f = L\pi d_f = S_f$ представляет собой общую поверхность волокон, т.е.

$$f = \frac{S_f}{S} \cdot \frac{1}{\pi(1-\alpha)} \tag{8}$$

Общую поверхность волокон S_f можно определить через удельную поверхность S_{ud} [м^2/кг] фильтрующего материала следующим образом: $S_f = S_{ud} \cdot m$, где m – масса фильтрующего материала, а определяется, например, методом БЭТ.

Полагая $1 - \alpha \approx 1$, получим упрощенную формулу для определения f:

$$f = \frac{S_{ud} m}{S\pi} \tag{9}$$

Таким образом, в основном уравнении фильтрации (2) с учетом (8) структура материала в общем случае характеризуется двумя параметрами: поверхностью волокон и плотностью упаковки. Из уравнения видно, что чем выше суммарная поверхность волокон и плотность

упаковки, тем выше задерживающая способность фильтрующего материала (меньше коэффициент проскока). Это хорошо согласуется с экспериментальными данными (см., например [2]) и достаточно ясно объясняет, почему в процессе накопления пыли на волокнах фильтра повышается эффективность фильтрации. Основной проблемой при решении уравнения (2) является определение суммарного коэффициента захвата волокон единичной длины ξ'_{Σ}.

Для грубых оценок эффективности в (2) вместо ξ'_{Σ} можно подставлять значение эффективности осаждения на одиночном цилиндре ξ_{Σ}, однако величины этих параметров имеют значительное расхождение [3], связанное с тем, что поле скоростей при обтекании волокон в фильтре значительно сложнее, чем при обтекании одиночного цилиндра, принятого в качестве модели.

При известном коэффициенте захвата одиночным цилиндром ξ суммарный коэффициент захвата фильтра или его компоненты с учетом величины α можно грубо оценить следующими зависимостями [3]:

$$\xi_{\Sigma} = \xi(1+4{,}5\alpha);\ \xi_{\Sigma}^{Stk} = \xi(1+110\alpha);\ \xi_{\Sigma}^{D} = \xi_{D}(1-4\alpha);\ \xi_{\Sigma}^{R} = \xi_{R}(1+30\alpha). \qquad (10)$$

В высокоэффективных волокновых фильтрах волокна обычно очень тонкие и полидисперсные. Они ориентированы случайным образом и неравномерно распределены по объему фильтрующего материала. В связи с этим, уравнения для определения эффективности осаждения (кольматации), полученные на базе ячеистых моделей, не являются в полной мере адекватными и пригодны лишь для оценочных расчетов. Поэтому Н.А. Фуксом с сотрудниками [4] была предложена так называемая веерная модель, представляющая собой систему рядов, параллельных цилиндров, каждый предыдущий ряд, в которой повернут относительно предыдущего на некоторый угол. Веерная модель наиболее полно соответствует реальным условиям фильтрации. При одинаковых с реальным фильтром плотностью упаковки α и диаметром волокон d_f сопротивление такой упорядоченной структуры максимально [5].

В соответствии с теорией Кирша – Стечкиной − Фукса [6, 7] гидродинамический фактор для веерной модели фильтра, с учетом эффекта скольжения газа около тонких волокон, записывается следующим образом:

$$K_0^{\,b} = -0{,}5\ln\alpha - 5{,}52 + 0{,}64\alpha + \tau'(1-\alpha)K_n \qquad (11)$$

где $K_n = \dfrac{2\lambda}{d_f}$ – число Кнудсена, τ' – численный коэффициент, для веерной модели равный 1,43.

Неоднородность структуры в реальных фильтрах влияет на суммарный коэффициент захвата и на сопротивление фильтра. Для учета неоднородности введен коэффициент неоднородности структуры $E^{в}$, характеризующий отношение силы, действующей на единицу длины волокна в верной модели $F_{\Phi}^{\hat{a}}$ к силе, действующей в реальном фильтре F_{Φ}^{p}:

$$E^{в} = \frac{F_{\Phi}^{в}}{F_{\Phi}^{p}} = \frac{\xi'^{\,в}_{\Sigma}}{\xi'^{\,p}_{\Sigma}} \qquad (12)$$

В реальных фильтрах волокна полидисперсны. Мерой полидисперсности является квадратичное отклонение σ диаметров волокон от среднего

$$\sigma = \frac{\overline{d_f}^2 - \overline{d_{f,i}}^2}{\overline{d_{f,i}}^2} \tag{13}$$

Тогда, гидродинамический фактор с учетом полидисперсности волокон, запишется:

$$K_0^в = -0,5\ln\left(\frac{\alpha}{1+\sigma}\right) - 0,52 + 0,64\left(\frac{\alpha}{1+\sigma}\right) + \tau'(1-\alpha)K_n \tag{14}$$

.

При известном экспериментальном значении сопротивления фильтра $\Delta P_{экс}$ можно определить $F_ф^p$ по формуле:

$$F_Ф^p = \frac{\Delta P_{экс}\pi^2 d_{f,i}^{-2}(1+\sigma)}{4\alpha\mu U_0 H} \tag{15а}$$

а силу сопротивления в веерной модели

$$F_ф^в = \frac{4\pi}{K_0^в} = \frac{4\pi}{-0,5\ln\left(\frac{\alpha}{1+\sigma}\right) - 0,52 + 0,64\left(\frac{\alpha}{1+\sigma}\right) + \tau'(1-\alpha)K_n} \tag{15б}$$

Коэффициент проскока в реальном фильтре для определенного значения U_0, рассчитывается по формуле, аналогичной (2)

$$K = \exp\left[-\frac{4\xi_\Sigma^{'в}\alpha H}{\pi d_f E^в (1+\sigma)}\right], \tag{16}$$

а эффективность фильтра

$$\xi_Ф = 1 - \exp\left[-\frac{4\xi_\Sigma^{'в}\alpha H}{\pi d_f E^в (1+\sigma)}\right] \tag{17}$$

По теории Кирша – Стечкиной − Фукса, являющейся по мнению многих авторов наиболее точной для модельных фильтров в области с доминирующими диффузией и захватом (например, [3, 8]), в области максимального проскока коэффициент захвата отдельным волокном определяется как

$$\xi_\Sigma^{'в} = \xi_D^в + \xi_R^в + \xi_{DR}^в \text{ или } \xi_\Sigma^{'в} = 1 - (1+\xi_D^в)(1-\xi_R^в), \tag{18}$$

где:

$$\xi_D^в = 2,7Pe^{-2/3}\left[1 + 0,39(K_0^в)^{-1/3}Pe^{1/3}K_n\right] + 0,62Pe^{-1};$$
$$\xi_R^в = (2K_0^в)^{-1}\left[(1+R_K)^{-1} - (1+R_K) + 2(1+R_K)\ln(1+R_K) + +2\tau' K_n(2+R_K)R_K(1+R_K)^{-1}\right]; \tag{19}$$
$$\xi_{DR}^в = 1,24(K_0^в)^{-1/2}Pe^{-1/2}R_K{}^{2/3}.$$

Рассчитать эффективность фильтра по (17) можно при условии, что диаметры волокон различаются не очень сильно (например, если $d_{f_1} < d_{f2}$, то теория верна при $d_{f2}/d_{f_1} \geq 0{,}2$). Теория верна также при $\mathrm{Re} \ll 1$, $\Delta P/U_0 = const$, $d_f \leq 1$ *мкм*, α<0,1.

Анализ приведенных выражений для определения ξ'_{Σ} показывает, что их в основном отличает друг от друга форма описания поля потока, т.е. гидродинамический фактор, зависящий от параметров модели фильтра, а также учет неоднородности структуры и полидисперсности волокон. Все корреляции не учитывают явно Re и применимы при следующих условиях: $\mathrm{Re}<1$, $\alpha<0{,}1$. В области чисто диффузионного осаждения (d_p<*0,3* мкм и U_0<*1* м/с) для волокон размером $d_f = 2$ мкм хорошее совпадение с экспериментальными данными дают корреляции Ли и Лиу на основе гидродинамического фактора Кувабары. Для фильтров из ультрадисперсных волокон (d_f< 1 мкм) более точными являются уравнения Кирша – Стечкиной – Фукса, полученные для веерной модели (19), которые учитывают неоднородность реального фильтра и полидисперсность волокон, а также эффект скольжения молекул газа около тонких волокон и совместное действие механизмов осаждения в результате диффузии и касания. Область применимости (19): $\mathrm{Re} \ll 1$, $d_f \leq 1$ *мкм*, α<0,1, $\Delta P/U_0 = const$, не очень большая однородность волокон, т.е. при $d_{f1} > d_{f2}$ должно выполняться условие $d_{f2}/d_{f_1} \geq 0{,}2$.

Коэффициент проскока и эффективность очистки газа модельным волокновым фильтром определяется соответственно из выражений (2) и (3), а реального фильтра по теории Кирша–Стечкиной–Фукса с учетом неоднородности структуры и полидисперсности волокон – из (16) и (17). Все эти и другие известные выражения аналогичны и содержат параметр фильтрации (3), который наглядно описывается соотношением общей поверхности волокон S_f фильтра к площади фильтра S, которое можно назвать удельной площадью волокон F^*:

так как

$$\alpha = \frac{\textit{объем волокон}}{\textit{объем фильтра}} = \frac{L\pi {d_f}^2}{4SH}, \text{ то } L = \frac{4\alpha HS}{\pi {d_f}^2}, \text{ и}$$

$$F^* = \frac{S_f}{S} = \frac{4\pi d_f}{S} = \frac{4\alpha H \pi d_f S}{\pi {d_f}^2 S} \qquad (20)$$

(этот результат соответствует и (5)), а параметр f определяется как

$$f = F^* \frac{1}{\pi(1-\alpha)} = \frac{F^*}{\pi\Pi}. \qquad (21)$$

При диффузионном осаждении, которое является доминирующим механизмом в области размеров частиц d_p < 0,5 мкм, осаждение происходит фактически только на 2/3 общей поверхности волокон. Если допустить, что действуют и другие механизмы задержки, то можно считать рабочей всю лобовую поверхность волокна плюс часть задней поверхности (по направлению потока), т.е. 5/6 общей поверхностью. Введение такой коррекции в формулу для определения F^* (21) вряд ли приведет к недооценке общей эффективности фильтра, т.к. данные теоретических расчетов, как правило, являются завышенными.

Выше отмечено, что общую поверхность волокон удобно, с точки зрения ее измерения, выразить через удельную поверхность: $S_f = S_{ud} \cdot m$, тогда

$$F_u^* = \frac{S_{ud} \cdot m}{S} F^* \quad \text{и} \quad f = \frac{F_u^*}{\pi \Pi} \tag{22}$$

или с корреляцией рабочей поверхности:

$$f = \frac{5}{6} \frac{F_u^*}{\pi \Pi}. \tag{23}$$

Для очень низкой плотности упаковки можно использовать упрощенную форму (9).

Учитывая вышеизложенное, можно записать:

– для однородного модельного фильтра:

$$\xi_\Phi = 1 - \exp(-\frac{5}{6} \frac{F_u^* \xi_\Sigma'}{\pi \Pi}); \tag{24}$$

– для однородного фильтра с учетом режима течения (из корреляции для коэффициента массопереноса):

$$\xi_\Phi = 1 - \exp(-\frac{5}{6} F_u^* \cdot 1{,}0664 \, \mathrm{Re}^{-0{,}489} N_{S_c}). \tag{25}$$

– для неоднородного фильтра из полидисперсных ультратонких волокон (на основе веерной модели):

$$\xi_\Phi = 1 - \exp(-\frac{5}{6} F_u^* \cdot 1{,}0664 \, \mathrm{Re}^{-0{,}489} N_{S_c}). \tag{26}$$

Параметр F_u^* (или F^*) может служить фактором оптимизации волокновых фильтров, поскольку учитывает общую поверхность осаждения, что удобно для определения пылеёмкости и ресурса работы фильтра, материалоемкость, геометрические и структурные параметры фильтра, и является безразмерным, что позволяет использовать его в сочетании с другими параметрами для сравнения фильтров между собой. Применение уравнений (24) − (26) для определения эффективности реальных фильтров требует учета неоднородности их структуры и полидисперсности волокна, которые можно описать известными формулами логарифмически нормального распределения.

Произведенная формализация процесса очистки газопылевых потоков в пористых волокновых материалах позволяет определить связи между переменными процесса в динамическом режиме и перейти к компьютерному моделированию для решения задача оптимизации.

Литература

[1]. Пилиневич, Л.П. Пористые порошковые материалы с анизотропной структурой: методы получения/ Л.П. Пилиневич, В.В.Мазюк, В.В.Савич, М.В.Тумилович. – Минск: Тонпик, 2006. – 268 с.

[2]. Хлебников, Ю.П. Фильтры систем кондиционирования воздуха и вентиляции/ Ю.П. Хлебников. – Москва.: Стройиздат, 1990. – 128 с.

[3]. Белоусов, В.В. Теоретические основы процессов газоочистки/ В.В. Белоусов – Москва: Металлургия. – 1988. – 256 с.

[4]. Фукс, Н.А. Механика аэрозолей./ Н.А. Фукс.– Москва: Изд-во АН СССР. – 1955. – 351 с.

[5]. Двухименный, В.А. Системы очистки воздуха от аэрозольных частиц на АЭС /В.А. Двухименный, Б.М. Столяров, С.С. Черный. – Москва: Энергоатомиздат, 1987. – 88 с.

[6]. Кирш, А.А. Эффективность аэрозольных фильтров, состоящих из ультратонких полидисперсных волокон / А.А., Кирш, И.Б., Стечкина, Н.А. Фукс // Коллоидный журнал, 1975. – №1. С. 41-46.

[7]. Kirch, A.A. The Theory of Aerosol Filtration with Fibrous Filters/ A.A. Kirch, I.B. Stechkina // In Fundamental of Aerosol Science / Ed. by David T. Show.: John Willy and Sohu, Inc., 1978. – P. 165-256.

[8]. Ужов, В.Н. Очистка промышленных газов от пыли. В.Н. Ужов, А. Ю. Вальдберг, Б. И. Мягков. – Москва: Химия, 1981. – 390 с.

КЛАСТЕРИЗАЦИЯ ПЛАЗМИД ПАЛОЧКОВИДНЫХ ФОРМ БАКТЕРИЙ И ИХ ВИДОВ С ИСПОЛЬЗОВАНИЕМ СПЕКТРОСКОПИИ

И.В. Кухарчук
Ассистент кафедры электронных вычислительных машин БГУИР

Д.И. Самаль
Заведующий кафедрой электронных вычислительных машин БГУИР, кандидат технических наук, доцент

Белорусский государственный университет информатики и радиоэлектроники, Республика Беларусь
E-mail: i.kukharchuk@bsuir.by, samal@bsuir.by

Abstract. . This paper describes the solution of the problem of classification for bacterial species and their nutrient media. The described algorithm allows classifying three species of bacteria using self-organized map, as well as the clustering of three nutrient media for E. coli bacteria using k-means. The paper presents the rationale for choosing a solution model, calculating its accuracy, and ways to increase the accuracy. The experiments show the proposed solution is capable of clustering spectrograms with an accuracy of 96%.

Возможность быстрой автоматической идентификации и/или классификации бактерий (их видов) в образце до сих пор является серьёзной проблемой в области микробиологии.

В рамках настоящей работы рассмотрены существующие подходы к проведению кластеризации содержимого штаммов бактерий на основе их спектрограммы, полученной при помощи настольного микроскопа комбинационного рассеяния.

Целью представляемого исследования являлась реализация алгоритма, самообучающегося в условиях ограниченного количества имеющихся спектрограмм; после обучения задачей алгоритма является последующая кластеризация входящих штаммов с необходимым уровнем точности, превышающим 90%. В рамках работы были решены следующие задачи:

- загрузки и подготовки данных (спектрограмм) к обработке;
- создания архитектуры совокупного алгоритма обучения и кластеризации видов;
- выбора алгоритма обучения для первичного разделения бактерий по видам;
- выбора алгоритма кластеризации для разделения бактерий по питательной среде, в которой они выращены.

Первым этапом проведённого исследования являлось формирование базы спектрограмм известных бактерий. Определение вида бактерий происходит на основе их содержимого, в частности – плазмидов, и их влияния на изменения в спектрограммах [1]. На данном этапе все спектрограммы сохраняются в виде сырых необработанных данных.

Классифицируемыми бактериями являются E. coli, shiwinella, lactobacilus. Спектрограмма с максимально различающимися интенстивностями названных бактерий представлены на рисунке 1. Однако, если рассматривать весь набор возможных спектрограмм, то различия между данными бактериями в большинстве случаев практически нивелируются. Тестовый набор каждого вида бактерий составляет 100 штаммов. Таким образом, исходная база данных составляла три набора по 100 штаммов для каждого из видов исследуемых бактерий, соответственно.

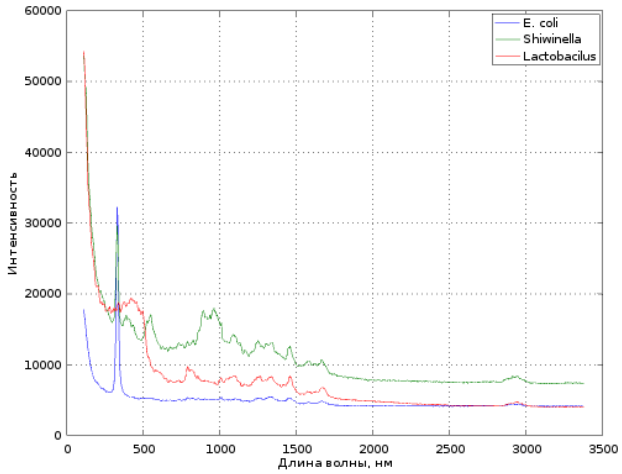


Рис. 1. Спектрограммы классифицируемых бактерий

Согласно постановке задачи, для первого классифицируемого вида бактерий – E. coli – необходимо различать бактерии по питательной среде, в которой они выращены (MMGls, LB и т.д.). Пример того, как среда влияет на изменение спектрограммы бактерии E. coli, отражён на рисунке 2.

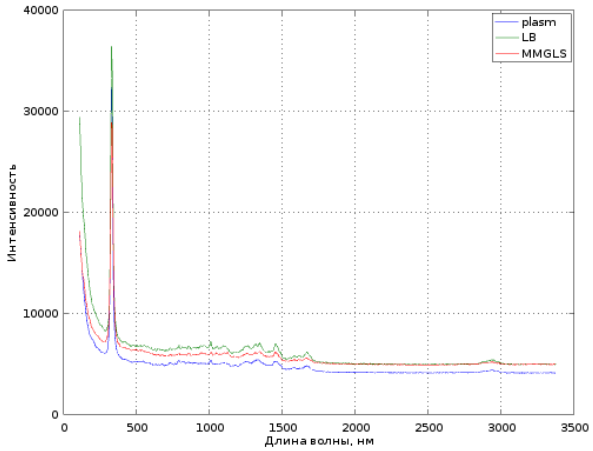


Рис. 2. Спектрограммы влияния различных питательных сред на примере бактерии E. coli

Так как имелась возможность использования тестового набора для классификации бактерий, то применялись алгоритмы с обучением, общий вид, которых, отображён на рисунке 3.

Для решения задачи классификации была выбрана нейронная сеть встречного распространения без обратных связей на базе самоорганизующейся карты Кохонена и звезды Гроссберга. Данная нейронная сеть хорошо подходит для указанной задачи ввиду следующих факторов:

– наличие свойства обобщения ввиду наличия слоя Кохонена;

– простота обучения слоя Гроссберга, накопления статистических данных;

– высокая скорость обучения [2-3].

Используемая в работе классическая схема нейронной сети отображена на рисунке 4. На данном рисунке входные данные представлены вектором X, слой Кохонена вектором K, слой Гроссберга вектором G, выходной вектор – Y. Веса для слоёв соответственно w и v.

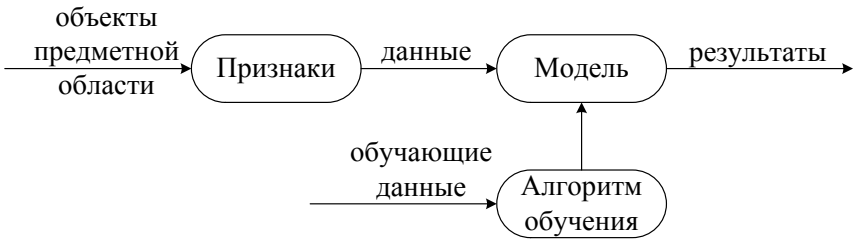


Рис. 3. Общий вид решения задачи машинного обучения

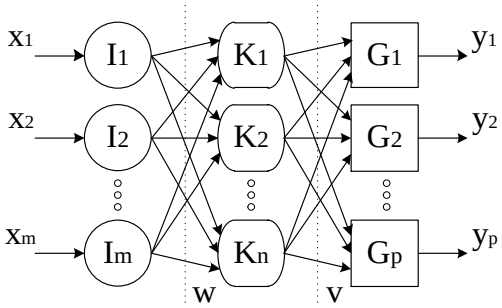


Рис. 4. Нейронная сеть с встречным распознаванием без обратных связей

После анализа сформированной базы спектрограмм в области влияния питательных сред на конечный вид графиков, была выявлена классическая закономерность: спектрограммы могут быть собраны в кластере, как это отражено на рисунке 2. Количество кластеров по постановке задачи заранее известно. Таким образом, достаточным решением для задачи кластеризации питательных сред бактерии E. coli является метод k-средних, имеющий вид [4]:

$$V = \sum_{i=1}^{k} \sum_{x_j \in S_i} (x_j - \mu_i)^2 \qquad (1)$$

где k – число кластеров; S_i – полученные кластеры; i = 1, 2, …, k и μ_i – центры масс векторов $x_j \in S_i$.

Архитектура решения, позволяющего классифицировать бактерии, и кластеризировать питательные среды для одной из них представлена на рисунке 5.

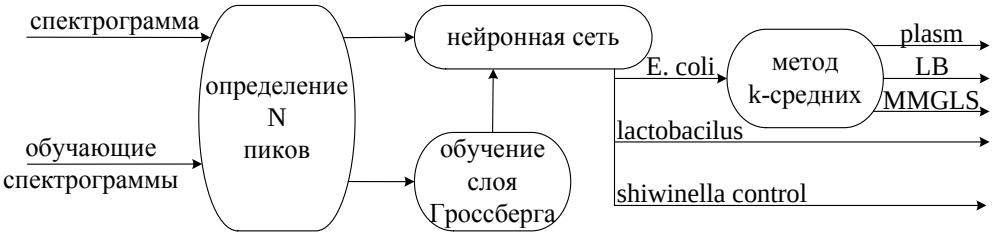


Рис. 5. Архитектура предложенного решения классификации бактерий и кластеризации их питательных сред

Входные данные построенной по образцу сети *могут* быть непрерывными, однако классификатор, работающий с таким объёмом непрерывных данных (1024 точки), продемонстрировал недостаточную точность определения бактерий (порядка 50%) на всех трёх типах. Для повышения точности было проведено редуцирование исходных данных и изменение исходных признаков, по которым проводилась классификация. В качестве признаков отбирались

пики спектрограмм и их общее количество для каждого объекта было сокращено до 100. Таким образом, операцией обработки входных данных нейронной сети стал отбор пиков. Для отбора определённого количества пиков изменялось разрешённое расстояние между ними. Данное действие позволило поднять общую точность определения бактерий до уровня 84%.

Следующими этапами модификаций нейронной сети стали:

– расширение обучающей выборки путём добавления шума к входным векторам;

– увеличение порога нейрона, который чаще остальных становится победителем;

– использование интерполяции вместо аккредитации– эмпирическая коррекция параметров α и β на этапе обучения нейронной сети.

В результате для каждого типа бактерии в классификаторе были получены изменения начального значения коэффициентов обучения следующим образом: . было создано поле в рамках $\alpha \in [0{,}41; 0{,}90]$ и $\beta \in [0{,}01; 0{,}50]$, и в результате циклов обучения получены статистические данные. Проведённые эксперименты отображены на рисунке 6. Алгоритм уменьшения коэффициентов обучения подчиняется показательному закону распределения.

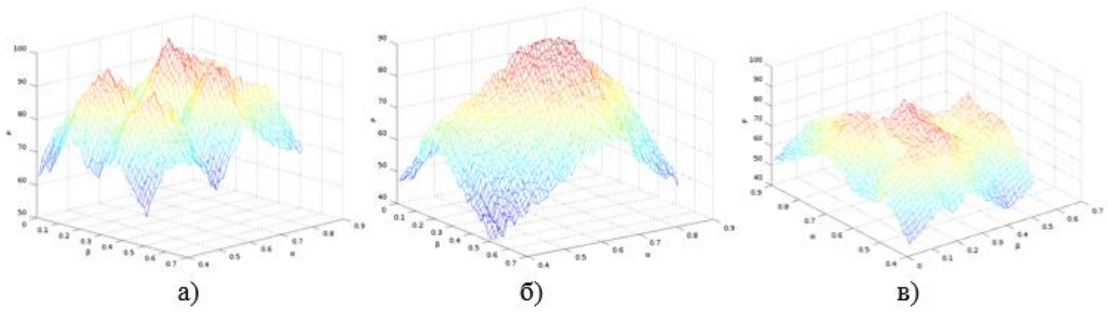


Рис. 6. Экспериментальные данные подбора коэффициентов обучения α и β нейронной сети встречного распространения без обратных связей: а) для бактерии E. coli, б) для бактерии shiwinella Control, в) для бактерии lactobacilus

На основании полученных данных выбраны следующие эмпирические коэффициенты: α = 0,85 и β = 0,15. Подбор начальных коэффициентов повысил точность модели на 2%. Совокупный эффект модификаций составил 12%.

В результате данной работы был спроектирован и реализован классификатор видов бактерий и последовательный кластеризатор питательных сред для бактерии E. coli с точностью классификации 96%.

Литература

[1]. Kong, K. Raman spectroscopy for medical diagnostics: from in-virto biofluid assays to in-vivo cancer detection / K. Kong, C. Kendall, N. Stone, I. Notingher // Advanced Drug Delivery Reviews. – 2015. – Vol.89. – Pp.121-134.

[2]. Olszewski, D. Time Series Visualization Using Asymmetric Self-Organizing Map / D. Olszewski, J. Kacprzyk, S. Zadrozny // Materials of 11th International Conference: Adaptive and Natural Computing Algorithms. – ICANNGA, 2013. – Pp.40-49.

[3]. Yu, D. Supervised Kernel Self-Organizing Map / D. Yu, J. Hu, X. Song, Y. Qi, Z. Tang // Materials of Third Sino-foreign-interchange Workshop: Intelligent Science and Intelligent Data Engineering. – IScIDE, 2012. – Pp.246-253.

[4]. Флах, П. Машинное обучение. Нака и искусство построения алгоритмов, которые извлекают знания из данных / П. Флах. – М.: ДМК Пресс, 2015. – 400с.

ПОСТРОЕНИЕ ГЕОСЕНСОРНЫХ СЕТЕЙ МОНИТОРИНГА ОКРУЖАЮЩЕЙ СРЕДЫ НА ОСНОВЕ INTERNET OF THINGS

К.С. Дик
Utech LLC location – Madison

И.С. Терех
Руководитель проектов, Theseus Lab s.r.o., кандидат технических наук

Е.А. Криштопова
Доцент кафедры инженерной психологии и эргономики БГУИР, кандидат технических наук, доцент

Utech Solution Inc, USA
Theseus Lab s.r.o., Czech Republic
Белорусский государственный университет информатики и радиоэлектроники, Республика Беларусь
E-mail: constantine.dzik@utechdata.com, it@theseuslab.cz, zam_po@mrk-bsuir.by

Abstract. The rapid growth of the number of user devices containing miniature sensors and having the ability to geolocation, laid the basis for the development of the deploying of geosensor networks. Such networks are capable of accumulating the huge amounts of data on various physical phenomena relative to the considered spatial and time intervals. Geosensor networks can provide the collecting and automatic processing of environmental data, allow to make the decisions to preserve it, predict natural disasters and ensure control over the consequences of technogenic disasters. The construction of geosensor network on hardware equipment of heterogeneous Internet of Things devices with their integration over the Internet using open OGC standards for data exchange about the state of the environment is considered.

Рост антропогенной нагрузки на окружающую среду приводит к ее существенному изменению. Современный мир характеризуется высоким риском техногенных катастроф и жесткими требованиями к их своевременному предупреждению и реагированию на природные стихийные бедствия.

Окружающая среда неоднородна и динамически изменяется, поэтому ее мониторинг требует высокого временного и пространственного разрешения. В течение последнего десятилетия феноменальный рост инфокоммуникационных технологий и технологий зондирования оказал существенное влияние на сферу наук о Земле.

Экологические проблемы имеют глобальный характер. Для эффективного понимания и мониторинга окружающей среды необходимы интегрированные данных из разнородных источников по всему миру. Другими словами, необходима открытая, масштабируемая, гибкая и устойчивая инфраструктура.

Современные модели миниатюрных недорогих датчиков с батарейным питанием, с поддержкой беспроводной цифровой связи представляют значительные возможности для решения задач наблюдения и управления. С ростом краудфандингового движения в мире, повсеместной цифровизации человеческого социума, популярности одноранговых сетей интересной видится возможность привлечения каждого отдельно пользователя к контролю параметров окружающей среды через его персональные коммуникационные устройства, оснащенные геолокационными и другими датчиками.

Геосенсор (геодатчик) - устройство, содержащее сенсоры (датчики) для снятия показаний с физического окружения с возможностью геопозиционирования точки снятия данных.

Геосенсорные сети (GeoSensor Networks - GSN) - это специализированные приложения технологий беспроводных сенсорных сетей в географическом пространстве, которые поддерживают геопозиционирование, наблюдают за изменениями и отслеживают перемещения явлений и процессов окружающей среды [1]. GSN могут обеспечить генерацию данных с высоким пространственным и временным разрешением, измерение разнообразия экологических данных и автоматизацию операций их обработки, повысить релевантность мониторинга окружающей среды. Учитывая эти особенности, GSN являются важной частью технологий, связанных с сенсорными сетями, а также многих новых концепций, в частности Internet of Things (IoT).

Интернет вещей (Internet of Things - IoT) — методология вычислительной сети физических предметов («вещей»), оснащённых встроенными технологиями для взаимодействия друг с другом или с внешней средой [2], рассматривающая организацию таких сетей как явление, способное перестроить экономические и общественные процессы, исключающее из части действий и операций необходимость участия человека [3].

Узлы со встроенными геосенсорами, разнородны с точки зрения аппаратных возможностей и протоколов связи, что усложняет возможность их участие в сценариях IoT. Поэтому обеспечение их совместимости является важным шагом для объединения различных устройств в геосенсорную сеть.

Учитывая эти трудности, международная некоммерческая организация по разработке стандартов в сфере геопространственных данных и сервисов Open Geospatial Consortium (OGC) [4], объединяющая 521 компанию, орган государственного управления и учебное заведение, предложила структуру открытых стандартов для использования подключенных к сети датчиков и сенсорных систем всех типов.

В данной работе обсудим, как сети геодатчиков могут быть интегрированы в Интернет, чтобы обеспечить задачи мониторинга окружающей среды, рассмотрим концепцию IoT и архитектуру слоев IoT для мониторинга экологических данных в режиме реального времени с использованием интгерированных в веб-сеть сенсоров.

Интернет вещей можно определить как динамическую глобальную сетевую инфраструктуру с возможностями самоконфигурирования на основе стандартных и совместимых протоколов связи, где физические и виртуальные «вещи» имеют идентификационные данные, физические и виртуальные атрибуты и используют интеллектуальные интерфейсы, а также легко интегрируются в инфокоммуникационные сети.

Благодаря использованию Интернет-интерфейсов и доступу к глобальной сети, IoT не ограничивается какими-либо физическими границами. Кроме того, его функционирование в режиме реального времени обеспечивает не только эффективный мониторинг окружающей среды, но также и возможности для международного сотрудничества, например, Глобальная система систем наблюдения Земли (ГЕОСС) и создание «умной Земли».

Внедрение парадигмы IoT в реальный мир требует новых технологий как на уровне сбора данных, так и на уровне сетевого взаимодействия. Наиболее часто используемыми технологиями на уровне сбора данных являются сенсорные сети, радиочастотная идентификация (RFID) и двумерные коды, Bluetooth, беспроводные локальные сети (WLAN) и Интернет - решения для сетевого взаимодействия. Сенсорные сети играют большую роль в большинстве приложений IoT, но гетерогенность в каждой сети с точки зрения аппаратных возможностей и протоколов связи влечет за собой использование множества разнообразных стандартов, например IEEE 802.15.4 и 6LowPAN [5]. В этом контексте консорциум OGC разработал инициативные стандарты встраиваемых в веб-сеть сенсоров Sensor Web Enablement (SWE), которые обеспечивают всестороннюю поддержку использования GSN в реализациях IoT [6]. Кодировки XML и сервисные интерфейсы для обнаружения, доступа и обмена любыми типами данных датчиков предложены в [7].

Разработанный OGC инициативный набор стандартов SWE обеспечивает совместимое

использование сенсорных ресурсов в унифицированном режиме. Эта инфраструктура позволяет локализировать, получать доступ, ставить задачи, а также оповещать о событиях в геосенсорной сети (Sensor Web). Таким образом, Sensor Web выступает в качестве инфраструктуры для сенсорных ресурсов, как и WWW для общих источников информации, что позволяет пользователям легко делиться своими геоинформационными ресурсами. Таким образом, SWE можно рассматривать как эффективное решение для сценариев IoT, связанных с сетями геодатчиков.

Модель интерфейса SWE состоит из следующих компонентов: служба наблюдения за датчиками (SOS), служба событий датчиков (SES), служба планирования датчиков (SPS) и служба веб-уведомлений (WNS). Наблюдения и измерения (O&M), язык моделей датчиков (SensorML), язык разметки датчиков (TML) относятся к информационной модели SWE, которая определяет модели данных главным образом для кодирования наблюдений датчиков и метаданных датчиков [8].

Наличие миллионов устройств, которые интегрируются или могут быть интегрированы в IoT, подразумевает необходимость стандартизации. Существует настоятельная потребность в адекватной архитектуре IoT, которая позволяет легко подключаться, управлять, обмениваться информацией и пользоваться полезными приложениями в рамках IoT. В настоящее время над стандартами для IoT работают независимо две рабочие группы крупнейших организаций по стандартизации – рабочая группа IEEE P2413 [9] и рабочая группа “Working Group on Internet of Things (WG10)” ISO/IEC [10].

Предлагаемая ISO/IEC [10] архитектура включает следующие уровни (рисунок 1):

Уровень устройства (Device Layer) отвечает за получение данных из физического мира, т.е. за свойства аппаратных устройств и свойства шлюза.

Сетевой уровень (Network Layer) отвечает за передачу данных. Фактически, этот слой отвечает за совместимость уровня устройства и прикладного уровня.

Уровень поддержки сервисов и поддержки приложений (Service support and Application Support Layer) отвечает за общие и специфические возможности поддержки сервисов и приложений.

Уровень приложения (Application Layer) предоставляет услуги или приложения для интеграции или анализа информации, полученной от предыдущих уровней.

Подобно WWW сенсорная сеть включает в себя три уровня – уровень данных (Data Layer), уровень веб-сервиса (Web Service Layer) и уровень приложений (Application Layer). Уровень данных в свою очередь может быть разделен на уровень физической среды (Physical Layer) и уровень сенсоров (Sensor Layer). Уровень данных обеспечивает наблюдение за параметрами окружающей среды и передает данные сенсоров уровню веб-сервиса. Уровень веб-сервиса обеспечивает доступ уровню приложения для извлечения кэшированных сенсором данных. Стековая структура сенсорной сети показана на рисунке 2 [11].

Системы управления сетью датчиков строятся на беспроводных сенсорных сетях (Wireless Sensor Networks - WSN) с использованием протоколов маршрутизации, оптимизации внутрисетевой связи и локализации датчиков в сети. Сенсорные веб-инфраструктуры обеспечивают доступ к ресурсам датчиков в Интернете и делают датчики доступными уровню приложений, создавая сенсорные веб-инфраструктуры. В этом классе некоторые соответствующие подходы используют стандарты SWE для обеспечения интероперабельного доступа к данным датчикам. Сенсорные веб-порталы обеспечивают доступность ресурсов сенсорных данных на уровне приложений и делают возможным для пользователей загружать и обмениваться данными с датчиками в нескольких форматах, например, числовых данных (например, измерения температуры), аудио- и видеоданных (например, данные веб-камер).

Архитектура геосенсорных сетей строится как онлайн платформа для сенсорной сети. Используя геосенсорные сети, пользователь может маневрировать сенсорным веб-браузером, используя виртуальную глобальную 3D или 2D карту, исследовать, визуализировать, получать

доступ и распространять гетерогенную и территориально распределенную информацию с сенсорных источников и другие связанные с ними данные. На рисунке 3 показана примерная архитектура геосенсорной сети GeoCENS (The Geospatial Cyberainfrastructure for Environmental Sensing) на базе инициативных стандартов OGC, которая обеспечивает упрощенный и эффективный поиск, публикацию, доступ к данным датчиков и совместное использование данных [13]. Подобно WWW, любой пользователь может построить и развернуть сенсорные веб-службы для размещения данных датчиков. Так как сенсорные веб-сервисы могут быть доступными по всему миру и не зарегистрированными ни в одном каталогизированном сервисе, очевидным становится необходимость поисковой машины для геосенсорной сети.

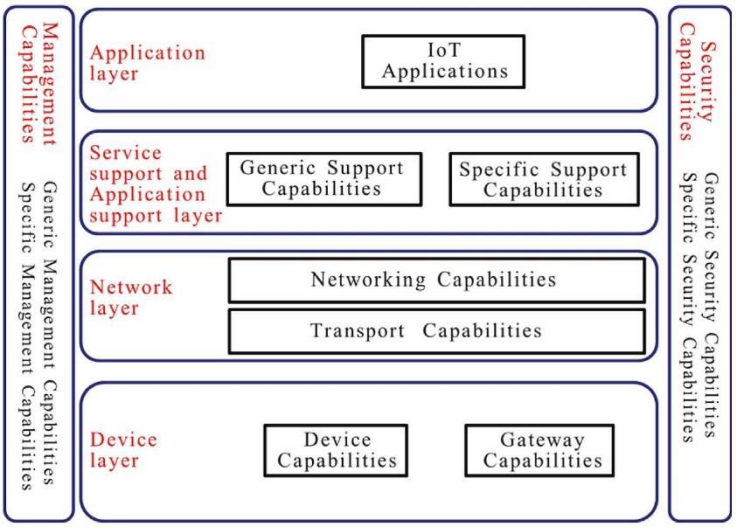


Рис. 1. Архитектура IoT согласно предложениям ISO/IEC [13]

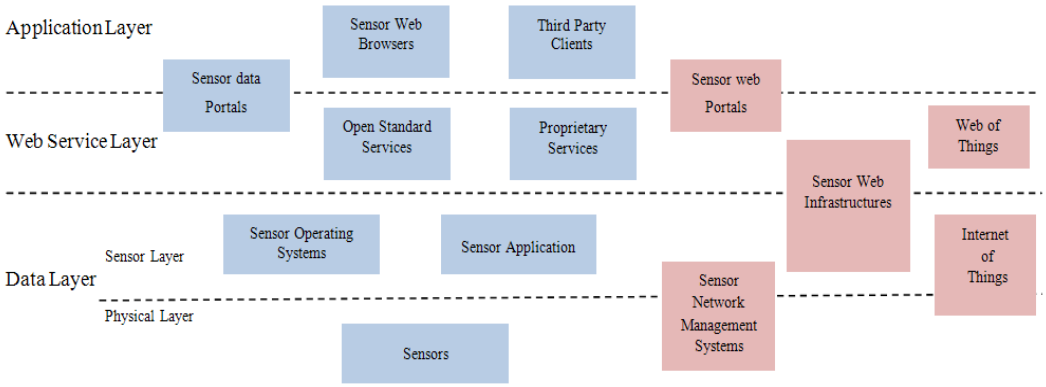


Рис. 2. Стек уровней сенсорной сети и место IoT в нем

Поскольку геодатчики оснащены различными средствами измерения параметров окружающей среды, например, температура, давление, влажность, скорость и направление ветра, а также наличие некоторых опасных загрязнителей (CO_x, NO_x, PM_x, SO_2), геосенсорные сети хорошо подходят для мониторинга окружающей среды, например для контроля загрязнения атмосферы. Сети геодатчиков с возможностью абсолютного или относительного позиционирования, а также сбора данных в режиме реального времени и точных данных с высоким пространственным и временным разрешением, имеют большой потенциал для предоставления геопространственной информации конечным пользователям.

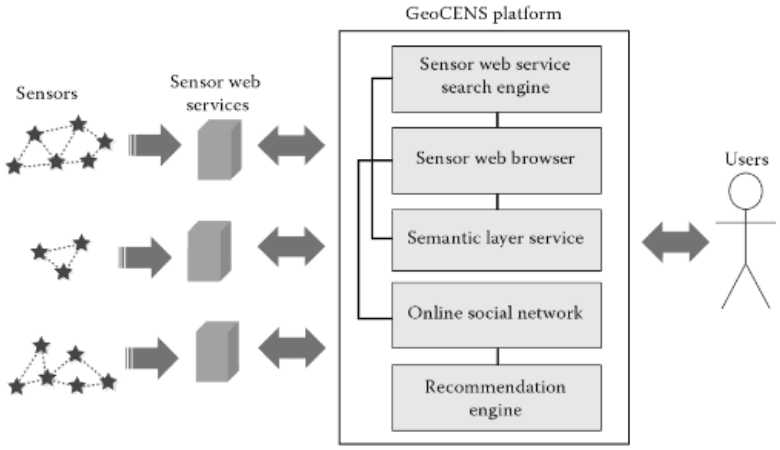


Рис. 3. Архитектура геосенсорной сети на базе платформы GeoCENS [13]

Узлы датчиков могут быть рассредоточены на большой территории и постоянно собирать данные об атмосферном загрязнении и другую необходимую информацию для определения состояния наблюдаемых точек и прогнозировать его тенденции на ближайшие время. Эта информация открывает новые возможности для лиц, ответственных за безопасное взаимодействие общества с окружающей средой.

Транспорт и промышленное производство являются основными причинами загрязнения воздуха в городской среде. Кроме того, некоторые природные условия, такие как топография, инверсия температур, влажность воздуха и ветер, могут усиливать или снижать концентрации загрязнителей воздуха. Таким образом, своевременное предоставление соответствующей информации об этих условиях, точность и надежность представления данных играют важную роль в борьбе с загрязнением воздуха, а также в мониторинге окружающей среды.

С помощью определенных методик анализа собранных датчиками данных реальным является точное по времени и местоположению прогнозирование стихийных бедствий (ураганы, землетрясения) и своевременное оповещение населения, в том числе в случае техногенных катастроф.

Интеграция геосенсорной сети с IoT, приведет к возможности доступа из любой точки мира к геопозиционированным данным об окружающей среде, что в свою очередь приведет к увеличению количества специализированных приложений, например мониторинга погоды в режиме реального времени.

Например, современные автомобили могут контролировать температуру воздуха и дороги с помощью встроенных датчиков. В настоящее время мобильные телефоны имеют GPS-датчик, акселерометр и компас, которые могут записывать аудио и видео из окружающей среды встроенным микрофоном и видеокамерами. Миниатюрные датчики давления, решения с двойным микрофоном для подавления внешнего шума и более специализированные датчики качества воздуха - все это уже начинает появляется в современных мобильных телефонах.

Низкая стоимость этих устройств и наличие их у огромного количества пользователей делает их пригодными для мониторинга окружающей среды. По сравнению с профессиональными системами они не обеспечивают высокой точности измерений, но могут дополнять их. Их количество и методы обработки данных на основе парадигмы Big Data дают возможность сделать данные пользовательских геосенсоров сопоставимыми с профессиональными и использовать в принятии решений.

Доступ в режиме реального времени или почти в режиме реального времени к данным, генерируемым любыми типами сетей геодатчиков, такими как станции мониторинга качества воздуха, смартфоны, камеры, биосенсоры или даже спутниковые изображения через Интернет, создает возможность построения «умной» системы контроля района, города, региона или

страны, в разрезе мониторинга окружающей среды, городского транспорта, здравоохранения или промышленности. Эта система поможет привести воздействие человека на окружающую среду в разумные рамки. Она также может рассматриваться как инструмент для решения глобальных проблем в рамках глобальной экосистемы групп и организаций, которые создают и используют данные наблюдений Земли (Global Earth Observation System of Systems - GEOSS) [14].

В рамках рассмотренной концепции IoT требуется, чтобы устройства получали данные измерения параметров физического мира. Устройства IoT неоднородны с точки зрения аппаратных возможностей и протоколов связи. Поэтому обеспечение интероперабельности является важным шагом для совместной интеграции различных устройств. Одним из решений этой проблемы является инициатива Sensor Web Enablement (SWE), которая представляет собой структуру открытых стандартов, обеспечивающих интероперабельное использование подключенных к веб-интерфейсу сенсорных ресурсов путем геопозиционирования, доступа, задач, а также событий и оповещений. Способность компонентов IoT для сбора данных открывает возможности создания широкого спектра приложений для мониторинга окружающей среды.

Литература

[1]. Nittel, S. A survey of geosensor networks: advances in dynamic environmental monitoring / Sensors Journal. - vol. 9, 2009. - P. 5664-5678.

[2]. Gartner IT glossary [Электронный ресурс]. – Режим доступа: http://www.gartner.com/it-glossary/ - Дата доступа 30.03.2017.

[3]. Gershenfeld, N., Krikorian R., Cohen D. The Internet of Things // Scientific American,. - Oct, 2004. – P. 76–81

[4]. The Open Geospatial Consortium (OGC) [Электронный ресурс]. – Режим доступа: http://www.opengeospatial.org/ogc/markets-technologies/swe - Дата доступа 30.03.2017.

[5]. Tan, J.; Koo, S.G. A Survey of Technologies in Internet of Things // IEEE International Conference on Distributed Computing in Sensor Systems, 2014. –P. 269.

[6]. Sensor Web Enablement (SWE) [Электронный ресурс]. – Режим доступа: http://www.opengeospatial.org/ogc/markets-technologies/swe - Дата доступа 30.03.2017.

[7]. Tamayo, A., Granell, C., Huerta, J. Using SWE standards for ubiquitous environmental sensing: a performance analysis // Sensors. – 2012. - 12(9). – P. 12026-12051.

[8]. Bröring A., Echterhoff J., Jirka S., Simonis I., Everding T., Stasch C., Lemmens R. New generation sensor web enablement / Sensors. – 2011. - 11(3). – P. 2652-2699.

[9]. Standard for an Architectural Framework for the Internet of Things (IoT) / IEEEE Standard Association [Электронный ресурс]. – Режим доступа: http://grouper.ieee.org/groups/2413/. - Дата доступа 30.03.2017.

[10]. ISO/IEC JTC 1 — Information Technology / International Organization for Standardization[Электронный ресурс]. – Режим доступа: https://www.iso.org/isoiec-jtc-1.html. - Дата доступа 30.03.2017.

[11]. Karimi H. A. Big Data: Techniques and Technologies in Geoinformatics. - CRC Press, 2014. – 312 c.

[12]. ISO/IEC JTC 1 - Information Technology / International Organization for Standardization [Электронный ресурс]. – Режим доступа: https://www.iso.org/isoiec-jtc-1.html - Дата доступа 30.03.2017.

[13]. Group On Earth Observations [Электронный ресурс]. – Режим доступа: http://www.earthobservations.org/geoss.php - Дата доступа 30.03.2017.

BIG DATA ДЛЯ ТРАНСПОРТНО-ЛОГИСТИЧЕСКИХ УЗЛОВ

А.А. Навроцкий
Заведующий кафедрой информационных технологий автоматизированных систем БГУИР, кандидат физико-математических наук, доцент

Р.В. Козарь
Студент кафедры информационных технологий автоматизированных систем БГУИР

Белорусский государственный университет информатики и радиоэлектроники, Республика Беларусь
E-mail: navrotsky@bsuir.by, pozitr0n.kozarroman@gmail.com

Abstract. Рассматриваются вопросы использования методов Big Data в системе 1С: Предприятие.

В настоящее время большинство предприятий занимаются решением вопросов, связанных с транспортной логистикой. Как правило, для построения своей логистической системы, предприятия пользуются услугами компаний, специализирующихся на решении задач по управлению процессами перевозок. Если предприятие предполагает использовать только собственную транспортную базу, не привлекая сторонних перевозчиков, оно может столкнуться с рядом проблем:

- изношенность подвижного состава;
- плохая информационная поддержка процесса транспортировки;
- сложности построения маршрутов перевозки;
- недогруз подвижного состава;
- необходимость страхования груза и транспортных средств;
- сложности при организации взаимодействия различных видов транспорта;
- недостатки имеющегося ПО для данной области.

В настоящее время на рынке имеется большое число программных продуктов для решения задач учета расходов предприятий. Однако большая часть из них является продуктами для ведения складского учета и плохо подходит для использования в транспортной логистике. Причем, даже использование такой мощной платформы, как 1С: Предприятие не всегда гарантирует необходимую наглядность по расходам и корректность их отображения в управленческом учете с разбиением, например, по рейсам. Имеющиеся на рынке иностранные программы, например Shipnet, не учитывают национальных особенностей, а поэтому их сфера применения ограничена.

Для решения задач транспортной логистики на предприятии был разработан программный модуль, который подключается как внешний модуль в конфигурацию 1С: Предприятие «Управление торговлей». Модуль обладает обширным кругом возможностей, основными среди которых являются:

- распределение заказов по нескольким курьерам/машинам;
- прокладка оптимального маршрута доставки по клиентам/поставщикам;
- ручная корректировка маршрута;
- выгрузка маршрута в навигатор;

– обработка документов по поступлению и реализации товаров и услуг;
– получение полной детализации по процессу доставки товаров.

Программный модуль используется в качестве рабочего стола для менеджера по приему заказов, что позволяет организовать быстрый подбор товара, формирование заказа и определение маршрута доставки.

Так как менеджер ежедневно обрабатывает сотни заказов, для нескольких тысяч товаров, то встает проблема передачи больших объемов данных через разработанный модуль. Это требует разработки специальных технологических инструментов и решений для проведения производительной обработки с учетом динамического роста объема исходных данных.

Установлено, что объемы данных, проходящие через модуль грузоперевозок, являются большими, так как присутствуют три основных признака Big Data:

volume – большое количество данных, независимо от масштаба доступных ресурсов для проведения их обработки (количество обрабатываемых заказов может достигать нескольких сотен в сутки);

variety – разнородность и слабая структурированность данных (заказы покупателей и поставщиков вообще не структурированы и разнородны);

velocity – необходимость быстрой обработки данных с предельно оперативной выдачи результата (клиенты ожидают быстрого и точного выполнения заказа).

При использовании 1С: Предприятие необходимость в обработке больших данных может возникнуть в следующих случаях:

− вычисление производных от больших данных (перезаполнение регистров при изменении первичных данных; проведение большого количества документов; заполнение новых реквизитов в больших таблицах);

− выгрузка, загрузка и конвертация данных (слияние (консолидация) баз, требующая переноса значительной части данных в другую базу; обмен данными вследствие обработки больших данных; восстановление испорченных данных из копии базы).

Для повышения эффективности обработки больших объемов данных в модуле было выполнено следующее:

1. *Отключение регламентных заданий в СУБД.* Выполнение регламентных заданий в СУБД может серьезно снижать производительность обработки из-за ожиданий на блокировках и очередях аппаратных ресурсов. Поэтому их лучше временно отключать, но желательно обеспечить их автоматическое включение после завершения обработки, в том числе аварийного.

2. *Оптимизация записи объекта.*

2.1. Минимизация ожидания на блокировках данных. У регистров отключается разделение итогов, а после больших многопоточных обработок пересчитываются итоги в периоды минимальной нагрузки. Конфигурация переводится на управляемые блокировки и используется версионный режим MS SQL (read_committed_snapshot). Для анализа ожиданий на блокировках данных используются центр управления производительностью (для управляемых блокировок и блокировок СУБД) и анализ технологического журнала из подсистемы «Инструменты разработчика» (для управляемых блокировок).

2.2 Запись в режиме загрузки. Если допустимо, то используется запись в режиме загрузки (Объект.ОбменДанными.Загрузка = Истина). В этом режиме должен выполняться очень незначительный процент кода обработчиков и подписок событий записи, а платформа отключает ряд своих внутренних обработчиков и потому выполняется меньше вычислений (проверка уникальности кодов и номеров объектов).

2.3. Отключение итогов регистров. Оправдано для больших обработок изменения регистров, однако в этом случае перестают работать виртуальные таблицы регистров и многие отчеты, поэтому необходимо обеспечить после успешного и аварийного завершения обратное включение итогов. Так как после включения итогов они могут пересчитываться платформой

некорректно, то желательно делать полный их пересчет.

2.4. Отключение авторегистрации изменений. В распределенных базах допускается в случае, если обработка выполняется одновременно во всех базах и есть уверенность, что изменения объектов общих данных будут одинаковы во всех базах.

2.5. Отключение RLS. По-возможности, следует выполнять запись без использования механизма RLS (ограничения доступа к данным на уровне записей), так как запросы RLS могут служить причиной серьезных потерь производительности. В качестве вариантов отключения данных механизмов можно использовать привилегированный режим, так же можно использовать набор ролей, устанавливающий пустые RLS на нужных таблицах.

3. *Использование многопоточности.* Относительная скорость наращивания однопоточной производительности аппаратных ресурсов постоянно падает по технологическим причинам. Поэтому аппаратные ресурсы все больше растут в сторону многопоточной производительности, например, растет количество ядер в процессорах. Использование многопоточности усложняет программу, но делает ее более масштабируемой по скорости работы.

3.1. Разбиение на порции. Для многопоточной обработки необходимо распределить набор объектов между потоками. Обработка не должна быть чувствительной к порядку обработки порций (обработка любой порции не зависит от успеха обработки других). Нечувствительными к порядку обработки порций могут быть процессы загрузки-выгрузки данных, объединение (замена) дублей, свертка регистра, восстановление последовательности (документов) по разным комбинациям значений измерений, универсальная обработка объектов, проведение документов, не использующих результаты проведения других документов, заполнение реквизитов, пометка удаления и удаление. Примером чувствительных к порядку порций обработок может быть восстановление последовательности (документов) по одной комбинации значений измерений или проведение документов, использующих результаты проведения других документов.

3.2. Многопоточный обмен данными. При переводе в многопоточный режим усложняется логика процессов передачи данных (рис. 1).

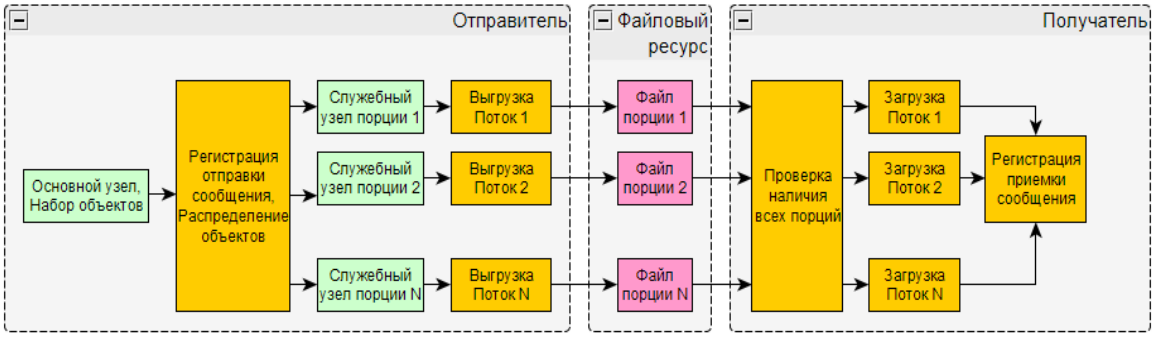


Рис 1. Логика процессов передачи данных

При оценке степени увеличения производительности следует учитывать:

– совокупную многопоточную производительности системы. Для ее оценки можно использовать Многопоточное тестирование производительности сервера 1С – СУБД;

– накладные расходы на многопоточность (построение карты порций, слияние результатов и др.);

– время ожидания при использовании блокировки данных (желательно использовать режим управляемых блокировок и разделения итогов регистров);

– количества потоков.

Многопоточное ускорение обработки набора объектов можно выразить формулой 1:

$$A = \frac{P + M \cdot T}{P + G + E + M \cdot (R + W \cdot T) + (1 - W) \cdot \frac{M \cdot T}{N}} \quad (1)$$

где, A – коэффициент ускорения (отношение длительности выполнения однопоточной обработки к длительности выполнения многопоточной обработки);

P − длительность не распараллеливаемой части вычислений для набора объектов в целом, общей для обоих вариантов обработки (например, выполнение сложного запроса для получения ключей объектов набора);

T − длительность вычислений на один объект в однопоточном режиме;

M − количество объектов;

G − длительность вычислений для набора объектов в целом, необходимых только для многопоточного варианта обработки (например, слияние результатов);

N − количество потоков;

E − длительность вычислений для порции объектов, необходимых только для многопоточного варианта обработки (например, сохранение результата порции);

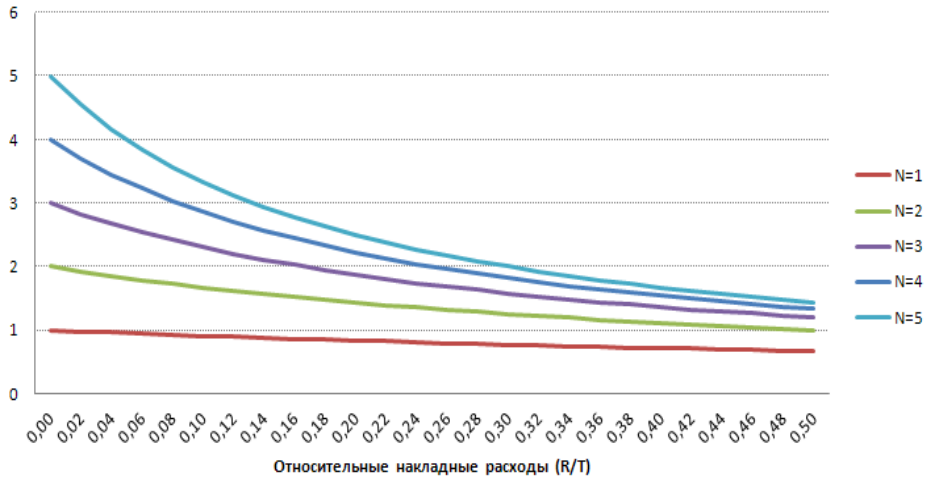
W − степень конкуренции (ожиданий, обусловленных многопоточным режимом), находится в диапазоне [0;1] и зависит от N, (основные типы ожиданий − на блокировках данных между потоками и на очередях аппаратных ресурсов);

R − длительность дополнительных вычислений по объекту в многопоточном режиме (например, привязка и отвязка объекта от порции)/

При достаточно большом количестве объектов и отсутствии конкуренции между потоками формула ускорения выглядит следующим образом:

$$\downarrow A = \frac{T}{R + \frac{T}{N}} = \frac{1}{\frac{R}{T} + \frac{1}{N}} \quad (2)$$

Таким образом оценив значения R и T, можно достаточно точно оценить ускорение для многих обработок. Чем меньше будут относительные накладные расходы (R по отношению к T), тем больше будет эффект от увеличения количества потоков.



Из рис. 2. видно, что однопоточный режим практически всегда быстрее многопоточного с количеством потоков (N = 1), что обусловлено необходимостью выполнения дополнительных вычислений. Поэтому при N = 1 многопоточный режим не эффективен.

С увеличением относительных накладных расходов уменьшается эффективность многопоточной обработки.

Литература

[1]. Фирма 1С-Битрикс. Подходы и инструменты работы с Big Data – все только начинается [https://habrahabr.ru/company/bitrix/blog/256551/] / Фирма 1С-Битрикс. – Электрон. текстовые дан. – Режим доступа https://habrahabr.ru/company/bitrix/blog/256551/, свободный.

[2]. Фирма 1С-Битрикс. Архитектура и технологические подходы к обработке Big Data на примере «1С-Битрикс BigData: Персонализация» [https://habrahabr.ru/company/bitrix/blog/272041/] / Фирма 1С-Битрикс. – Электрон. текстовые дан. – Режим доступа https://habrahabr.ru/company/bitrix/blog/272041/, свободный.

[3]. Microsoft Inc., 7 Things You Must Know About Big Data Before Adoption [http://bigdataanalyticsnews.com/7-things-must-know-big-data-adoption/] / Microsoft Inc. – Электрон. текстовые дан. – Режим доступа http://bigdataanalyticsnews.com/7-things-must-know-big-data-adoption/, свободный.

[4]. 1С-Прогресс, Многопоточная обработка данных в системе 1С: Предприятие [http://1sprogress.ru/mnogopotochnaya-obrabotka-dannyx-v-1s.html] / 1С-Прогресс. – Электрон. текстовые дан. – Режим доступа http://1sprogress.ru/mnogopotochnaya-obrabotka-dannyx-v-1s.html, свободный.

[5]. Фарит Насипов (infostart.ru), Как ускорить 1С – Многопоточная обработка данных [http://infostart.ru/public/306865/] / Фарит Насипов (infostart.ru). – Электрон. текстовые дан. – Режим доступа http://infostart.ru/public/306865/, свободный.

ТЕХНОЛОГИИ БОЛЬШИХ ДАННЫХ В РАБОТЕ С ПСИХОФИЗИОЛОГИЧЕСКИМИ ХАРАКТЕРИСТИКАМИ ПЕРСОНАЛА ЖЕЛЕЗНЫХ ДОРОГ И ВОДИТЕЛЕЙ АВТОМОБИЛЬНОГО ТРАНСПОРТА

Н.В. Камкичёва
Ведущий специалист Белорусской железной дороги, магистрантка кафедры инженерной психологии и эргономики БГУИР

Г.А. Розум
Ассистент кафедры инженерной психологии и эргономики БГУИР, магистр техники и технологии

В.В. Савченко
Директор Научно-инжинирингового центра «Бортовые системы управления мобильных машин», доцент кафедры инженерной психологии и эргономики БГУИР, кандидат технических наук

Н.В. Щербина
Старший преподаватель кафедры инженерной психологии и эргономики БГУИР, магистр технических наук, аспирант БГУИР

К.Д. Яшин
Заведующий кафедрой инженерной психологии и эргономики БГУИР, кандидат технических наук, доцент

Белорусский государственный университет информатики и радиоэлектроники, Республика Беларусь
E-mail: kafipie@bsuir.by, rozum@bsuir.by, uus@tut.by, shcherbina@bsuir.by, yashin@bsuir.by

Abstract. Psychophysiological qualities of drivers of vehicles have a decisive impact on road safety. Currently, there is a large number of technical tools used to study psychophysiological functions and characteristics with the purpose of conducting professional selection for various types of transportation. The development of relaxation skills in healthy people using systems with biological feedback was studied. Studies of professionally important qualities in the professional selection of drivers of vehicles have been carried out.

Введение. Тоническая и фазическая составляющие электродермальной активности человека позволяют судить о преобладании симпатических или парасимпатических процессов в организме человека в момент исследования [1]. Использование биоадаптивной компьютерной программы NeuroDog (разработка компании «Нейроком», Россия, версия 25.11.2013) может позволить научить человека навыкам самостоятельной активации механизма релаксации.

Для поддержания оптимальной работоспособности персонала железной дороги, непосредственно обеспечивающего перевозочный процесс (машинисты, помощники машинистов)

имеются санаторно-профилактические учреждения. Психологи проводят индивидуальные занятия по управлению психофизиологическим состоянием работников. Одним из основных направлений мероприятий, проводимых с машинистами и помощниками машинистов, является профилактика стресса.

Обеспечение безопасности движения является важнейшей задачей не только на железных, но и на автомобильных дорогах. В современных условиях управление транспортным средством осложняется высокой интенсивностью дорожного движения и участием в нем водителей с различной профессиональной подготовкой. Эти обстоятельства обуславливают значительное возрастание психических нагрузок и существенно повышают вероятность развития негативных изменений психофизиологического функционального состояния человека.

Цель настоящей работы – исследовать психофизиологические характеристики здоровых людей и определить подходы обработки полученных разноплановых экспериментальных данных.

Задачи работы следующие: 1) Рассмотреть теоретические и практические вопросы исследований психофизиологических качеств здоровых людей. 2) Исследовать развитие навыков релаксации здоровых людей (машинистов железнодорожных локомотивов). 3) Исследовать профессионально важные качества водителей транспортных средств.

Развитие навыков релаксации у здоровых людей с использованием системы с биологической обратной связью. Профессиональная психологическая работа с машинистами и помощниками машинистов на железной дороге ведется в нескольких направлениях: тестирование при приеме на работу; профессиональные плановые осмотры психолога по графику в зависимости от группы профессиональной годности; внеплановые осмотры (после травмы, продолжительной болезни, создания аварийной ситуации и др.); вспомогательные психологические мероприятия (консультации по релаксации, активации, по преодолению стресса, по поддержке своего функционального состояния на оптимальном уровне и др.).

Ранее нами был подготовлен сравнительный анализ известных методов и систем для анализа известных методов и систем для развития навыков релаксации у здоровых людей [2, 3]. Далее было принято решение о проведении эксперимента по выработке навыков релаксации у железнодорожного персонала с использованием программно-аппаратного комплекса NeuroDog [4].

Цель исследований заключалась: 1) в проверке возможности выработки релаксации у испытуемых при использовании визуальной биологической обратной связи по параметрам электродермальной активности; 2) установлении среднего количества сеансов, необходимого для обучения навыкам релаксации; 3) определении индивидуальных взаимосвязей между возможностью и скоростью выработки навыка релаксации у испытуемого и психофизиологическими качествами, диагностируемыми при профессиональном психологическом осмотре испытуемых.

В таблице 1 приведен пример результатов исследований испытуемого №1 (фамилия закрыта). С – сеанс, его порядковый номер.

Таблица 1. Пример результатов исследования одного из машинистов (двадцать сеансов)

№	С	Время, с												
1	1	5,8	3,6	7,3	13,4	10,4	4	6,3	5	9,7	4	5,9	3,7	→
		9,8	5,3	5,8	4,7	4,9	5,4	13,7	9,9	7	7,3	5,5	8,3	→
		5,2	6,7	7,6	5,4	4,3	82	33,4	77,1	260	7,6	12,2	61,7	→
		5,6	21,1	17,8	24,4	85	4,2	94,7						
	2	10,9	8,2	5	4,6	7,3	3,5	8,8	10	3,2	4,6	3,2	5,6	→
		6,3	7,5	6	6,2	3,9	6,7	4,1	7,5	3,2	4,8	5,7	4,4	→
		23,8	4,5	4,7	57,5	9,4	3,2	13,6	4,7	5,2	5	9,2	5,4	→
		4,2	9,9	12,1	5,5	4,8	4,9	5,5	145	36,5	17,8	11	3,9	→
		8,7	52,3	5,8	17,2	8,4	140	77,2	5,3	19,3	5	47,1	9,3	→
		50,2	3,2	9,8	4,8	7,7	25,7	54,7						
	3	15,8	4,2	8,5	5	12,5	7,6	3	3,1	3,2	3,6	5,9	5,7	→
		7,1	5,7											
	4	3,5	5,6	9,5	10,1	6,3	3,8	7,1	7,4	4,8	7,2	7,6	4,5	→
		11	6,9	5,7	9,4	3,9	24,8	21,8	6,2	3,4	14	3,3	5,6	→
		16,3	16,6	8,2	21,1	4,8	3,6	15	4,2	5,7	5,4	22,8	8,8	→
		20,4	108	13,6	20,9	4,6	3,2	19,2	5,8	6,5	12,4	35,3	15,2	→
		10,9	4,2	4,5	7	3,9	10,7	12	3,1	3,7	3,8	16,5	7,6	→
		11	6,1	22,7	3,6	6	6,5	20	8,1	82	51,9	14,3	11,2	→
		127	50	21,3	7,2	8,2	18,2	10,4	17,1	128	3,3	31,3	5,5	→
		14,9												
38 лет	5	5,2	4,7	15	11,1	3,2	3,6	8,2	5,9	8,5	4,3	15,3	5,1	→
		25,5	40	3,8	6,1	34,5	4,4	8,2	7,6	5,1	4,5	30,1	69,4	→
		8,6	82	6,7	5,7	32,3	71,7	6,5	8,6	10,1	4,4	6	21,1	→
		44,4	62,3											
	6	10,5	3,6	3,7	3,6	8,1	5,8	5,6	3,2	6,6	19,6	9,7	46,8	→
		22,2	36,3	20,2	12,1	20	6,1	26,2	5,2	5,1	12,9	3,8	15,4	→
		222	3,5	13,1	26,5	38,3	23,3	14,2	18,7	6,7	7,4	5,4	14,1	→
		6,4	37,2	18,4	25,5	4,5	3,5	18,6	27,7	23,5	4,2	22,1	38,6	→
		3,8	5,9	10,2	6,8	28,2	4,8	10,4	13,7					
	7	3,2	16,7	31,7	4,1	5,5	34,4	11,3	33,6	4,1	41,6	6,6	4,3	→
		19,8	4,8	11,6	3,9	36,4	6	18,8	4,1	9	4,9	10	4,3	→
		4,1	4,5	3,3	8,2	5,5	5,4	5	14,1	4	19,2	4,8	5,4	→
		4,2	3,3	4,2	9,2	9,4	8,9	9,1	4,5	3,5	4,9	4,6	4,3	→
		3,9	13,4	12,7	5,4	4,8	3,3	5,9	7,6	5,3	9,7	8,2	6,2	→
		11,1	8,1	5,2	3,6	32,3	10	7,4	3,9	13,6	5,9	7,5	3,3	→
		36,8	8,8	6,6	5	10								
	8	4,8	6,7	3,5	3,7	3,5	6,5	4	11,3	7,5	18	23,3	16	→
		4,4	16	6,8	20,2	9,1	5,5	12	15,7	113	6,8	5	4,8	→
		13,5	7	3,6	3,1	16	5,7	4,8	27,4	6,1	7	11,1	6,8	→
		5,8	4	6,3	3,4	5,8	3,6	6,2	37,5	8,8	5,8	5,5	9,9	→
		10,2	4,4	22,3	14,2	3,7	3,9	30,1	37,8	5,1	30,9	10,1	20,2	→
		12,8	9,4	5,2	8	10,5	7,6	3,1	3,8	32,8	4,4	17,4	4,1	→
		4	4,2	3,8	3,9	4	13,8	4,1	7	9,2	4,5	11	4,7	→
		9,5	11	17,7	8,6	3,4	4,3	4,2	3,2	3,6	11,2	32,5	15,6	→
		13,4	12,3	3,3	5,5	3,8	7	8,6	4,9	26,5	73,7	7	31,5	→
		6,7	6	6,2	8	8,1	3,8	11,9	15	23,4	26	16,1	26,6	→
		11,2	3,9	12,2	24,5	41,6	4,5	42,1	26,7	16,9	51,5	60,8	4,3	

№	С	Время, с												
	9	51,3	429	7,6	4,2									
	10	52,5	24,7	7,9	16	21	23,4	8,8	65,6	3,3	317	14,2	8,2	→
		4,2	8,2	3,7	14,4	23,4	9,2	14,4	23,4	21,1	18,7	11	7,8	→
		18,5	9											
	11	3,7	11,1	3,6	6,8	7,2	6,9	9,1	6,7	10,3	6,1	11	11,8	→
		4,2	10,1	5,2	8	8,5	19,9	4,7	4,5	9,7	10,6	7	7,9	→
		3,3	5,3	3,3	3,3	5,9	6,6	3,2	3,5	5,2	3,6	5	15	→
		4,3	4,1	7,1	9,4									
	12	22,1	3,2	6,9	8,7	5,6	7,9	8,2	7,8	11,2	14,4	5,2	12,6	→
		6,1	6,9	10,7	4	3,1	6,2	4,1	5	5,2	3,7	7,9	21,5	→
		11,7	13,1	13,5	10,4	7,7	23,6	4,4	122	4,7	3,5	14,2	15	→
		3,2	5	9,5	5,2	5	3,7	6,8	3,5	5,2	19,7	11,9	6	→
		9	4,6	4,5	13,6	7,9	6	3,3	3,6	4,5	17,2	47,2	3,3	→
		61,3	59,6	20,9	35,2	31,6	21,2	5,5	10,3	11,7	4,9	27,1	4,5	
	13	5,9	4,8	16,5	7	12,7	24,9	49,2	168	83,6	49,1	43,3	38,3	→
		96,8	29	62,4	59,8	18,1	10							
	14	12,3	44	4,3	8,3	8,7	4,8	4,9	19,3	5	7,4	11,2	56,4	→
		3,6	11,2	4,9	4,9	6,7	24,8	11,7	5,1	17,3	3,1	6,7	11,7	→
		3	5,6	5,7	17,4	23,2	33	35,4	6,7	30,3	3,9	3,9	6,7	→
		9,4	13	96,2	37,2	8,2	101	7						
	15	5,5	3,3	7,9	3,4	16,8	51,2	26,6	26,2	14,3	25	5,1	4,9	→
		12,4	6	28,9	94,2	10,7	24	16,5	10,8	5,9	7,2	8,5	5,5	→
		102	3,4	6,2	25,6	4,9	25,5	14,9	11,2	16,7	3,6	3,6	10,4	→
		23,7	126	213	6,1	10,4	19,1							
	16	9,6	3,2	7,2	3,5	4,3	19,5	3,2	13,8	4,2	7,8	18	15,7	→
		7,8	59,6	24,6	11,1	6,8	6,2	5	11,5	4,7	12,1	10,9	6,8	→
		12,8	4,8	16,7	4,7	13,4	15,2	6,6	6,8	7,3	10,3	4,5	7,6	→
		8,3	16	3,3	6,9	4,5	4,2	15,5	7,7	4,4	12	11,7	16,7	
	17	40,7	9,9	3,4	6,2	21,4	13,5	3,5	5,8	3,1	10,9	14,3	9,6	→
		7,5	16,3	29	6	12,7	6	5,7	5,4	4,9	13,7	10,4	12,8	→
		6	3,6	28,4	11,1	9,2	14,4	9,8	6,9	4,8	7,6	19,8	28,5	→
		3,7	8,6	8,5	11,2	3,6	7,9							
	18	4,6	9	8	15	3,9	14,7	10,5	3,1	6,6	31,6	5,4	35,1	→
		3,9	11,2	7	3,2	3,6	13,3	27,2	21,6	9,4	10,4	13,8	29	→
		38,6	12,1	3,1	6,1	7,1	3,5	17,4	7,5	54,5	62,7	25,4	48,2	→
		42,6	166	161	27,1	112,7								
	19	30,3	5,2	32,3	7,2	4,1	3,6	18,3	29,8	10,1	13,2	12,1	4,1	→
		9,1	9	5,3	4,7	5,7	5	3,2	8,2	6,2	9,3	7,7	6	→
		4,8	5,8	7,7	8	5,3	5,7	3,4	3,7	4,8	4,1	5,1	4,5	→
		7	13,1	3,8	3,6	5,8	3,8	7,5	7,1	4,4	5,9	5	4	→
		3,3	7	5,9	5,9	5,6	6,7	9	3,6	3,3	7,1	10,2	5,7	→
		5,3	4,9	26,4	10,2	10,8	20,5	22,7	18	3,7	8,2	8,8	18,1	→
		4,5	6,7	5,9	19,4	12,7	3,1	11	12,9	7,6	3,4	11,9	11,8	→
		15,4	24,1	36,9	4,1	7,1	10,4	15	6,8					
	20	119	53,8	148	153,9	68	5,4	90	481	3,6				

После набора группы из 35-40 человек, положительно прошедших эксперимент, т.е. успешно достигших полной релаксации, проводится вторая часть этого эксперимента. Устанавливается индивидуальная взаимосвязь между результатами эксперимента по выработке навыков на релаксацию испытуемых и результатами, полученными при периодическом про-

фессиональном осмотре психологом. Здесь необходимо отметить следующее. При периодическом профессиональном осмотре машинистов и помощников машинистов электровозов психолог использует стандартизованный Комплекс универсальный психодиагностический УПДК-МК для профессионального психофизиологического отбора работников локомотивных бригад, диспетчеров.

Существуют зависимости (корреляции) между экспериментальными результатами, полученными при психофизиологическом осмотре испытуемого психологом с использованием УПДК-МК и экспериментальными результатами, полученными в результате использования экспериментального образца биоадаптивной компьютерной программы NeuroDog версии 25.11.2013. Т.е. существуют определенные индивидуальные психофизиологические параметры человека, которые влияют на скорость и возможность выработки им навыков на релаксацию (рисунок 1).

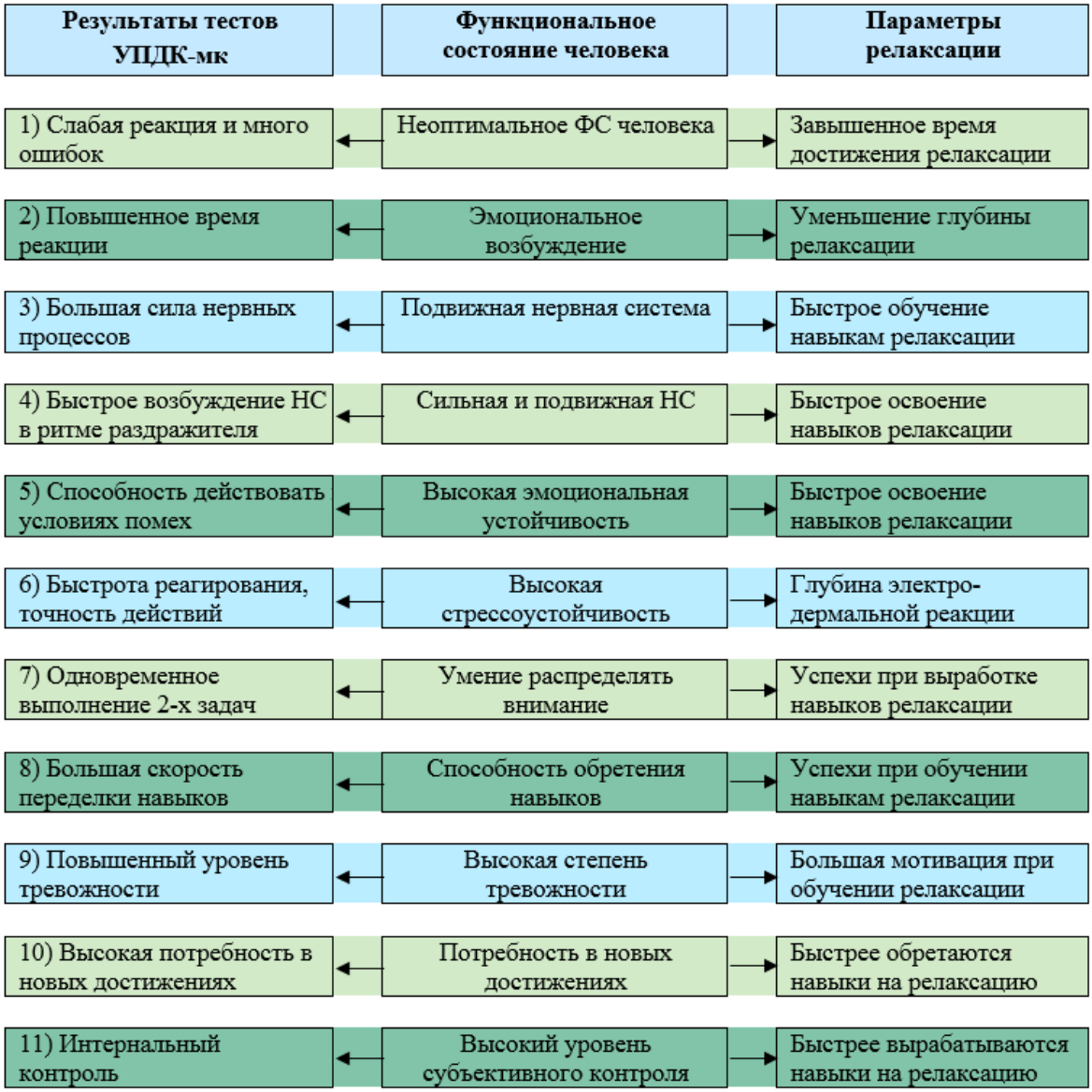


Рис. 1. Корреляция между результатами осмотра испытуемого психологом и исследованиями, полученными с использованием компьютерной программы NeuroDog

Такая связь существует для результатов, полученных при применении следующих методик из УПДК-МК (разработка компании «Нейроком», Россия).

1) Общая готовность к работе. Тест проводит динамический контроль функционального состояния оператора. При низком времени реагирования и большом числе ошибок испытуемый находится в неоптимальном функциональном состоянии, что влияет на увеличение времени, необходимого для его релаксации.

2) Реакция на движущийся предмет. Тест оценивает степень уравновешенности процессов возбуждения и торможения у испытуемых. Повышенное время реакции свидетельствуют об эмоциональном возбуждении, что влияет на глубину диапазонов релаксации. Время реакции в среднем находится в диапазоне 0,20 – 0,35 с.

3) Теппинг-тест. Позволяет выявить силу и подвижность процессов нервной системы. Человек быстрее обучается навыкам релаксации, если он обладает более подвижной нервной системы. Среди испытуемых нет людей со слабой нервной системой, поскольку испытания проводятся среди машинистов и помощников машинистов электровозов.

4) Критическая частота световых мельканий. Показывает наибольшую частоту световых мельканий, при которой нервная система человека возбуждается в ритме раздражителя. Люди с наиболее подвижной нервной системой быстрее осваивают навыки релаксации.

5) Эмоциональная устойчивость. Человек с высокой способностью действовать в условиях помех и негативных эмоциональных факторов способен быстрее освоить навыки релаксации.

6) Стрессоустойчивость. Умение мобилизоваться, сохранять точность и быстроту реагирования. Испытуемые с высокой стрессоустойчивостью имеют малый диапазон глубин электродермальной реакции.

7) Распределение внимания. Человек с высокой способностью одновременного выполнения двух задач также успешен и при выработке навыка на релаксацию: одновременное слежение за анимационной картиной и применение способов самоуспокоения.

8) Скорость переделки навыков. Работники с высокой способностью обретения навыков в похожих видах деятельности, успешны и при обучении навыкам релаксации.

9) Методика измерения уровня тревожности по Тейлору. Позволяет оценить уровень тревожности. При высокой тревожности человек имеет большую мотивацию к обучению навыкам на саморелаксацию. Возможно применение вместо этой методики теста Спилбергера по определению ситуативной тревоги и личной тревожности.

10) Потребность в достижениях. При средней и высокой потребности в достижениях человек быстрее обретает навыки на саморелаксацию.

11) Уровень субъективного контроля. Люди с интернальным контролем быстрее вырабатывают навык на саморелаксацию.

Исследование профессионально-важных качеств при профессиональном отборе. Сравнение категорий ошибок водителей транспортных средств с причинами ДТП показывает, что задержка распознаваемости предметов, в результате которой происходит около 50 % ДТП, зависит главным образом от временной ошибки. Принятие ошибочных решений водителями является причиной 40 % ДТП (ошибки исполнения). Важное значение для повышения надежности водителя имеет проверка состояния здоровья, функций органов чувств и изучение психофизиологических особенностей. Для такого изучения нужны соответствующие приборы и методы, обеспечивающие на должном уровне индивидуализированный подход при медицинском освидетельствовании водителей и кандидатов. Задачей исследований является анализ предпосылок ошибочных действий водителей, диагностика актуального состояния участников эксперимента, а также поиск методов коррекции уязвимых профессионально-важных качеств участников эксперимента – водителей и кандидатов.

На рисунке 2 представлены основные компоненты структуры деятельности водителя транспортных средств.

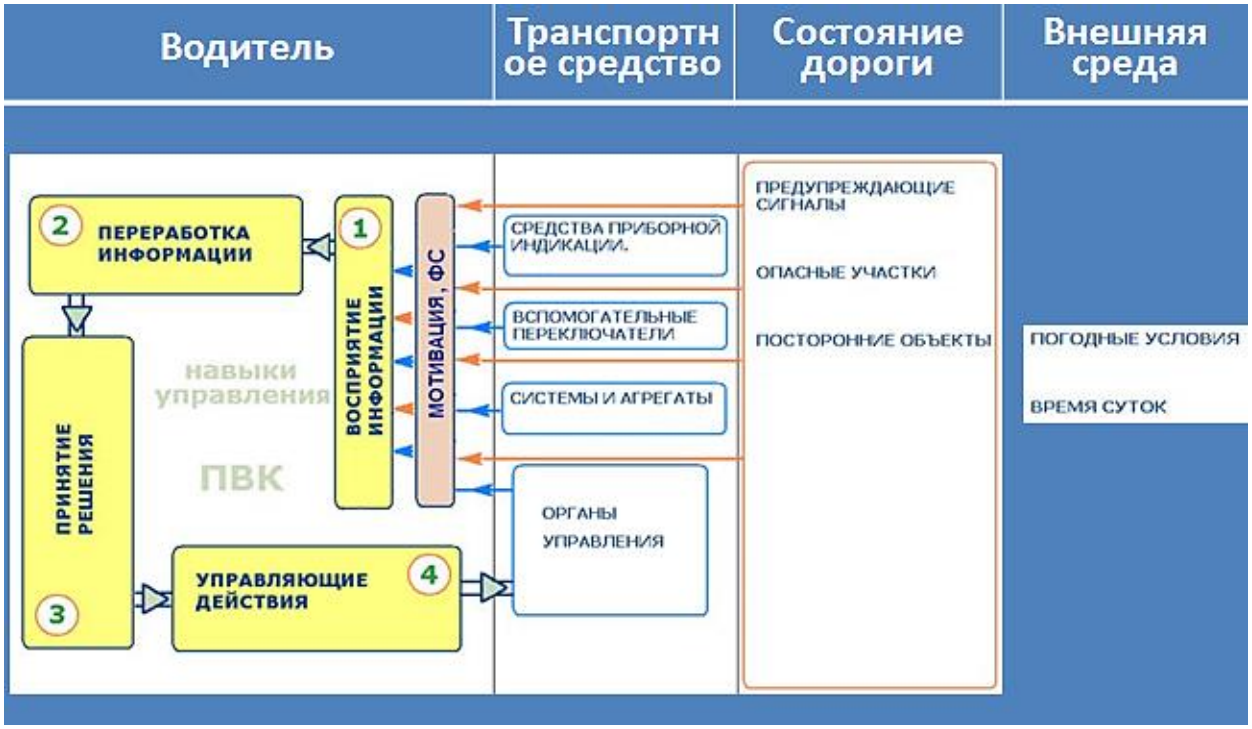


Рис. 2. Основные компоненты структуры деятельности водителя транспортных средств [5]

Применительно к транспортному процессу структурную схему системы эксплуатации транспортного средства можно представить состоящей из четырех основных элементов: водитель – автомобиль – дорога – среда [5]. Блок-схема компьютерных методов повышения уязвимых профессионально-важных качеств в составе мероприятий по профотбору водителей представлена на рисунке 3.

Рис. 3. Основные компоненты структуры деятельности водителя транспортных средств [5]

Для проведения экспериментальных исследований использованы аппаратно-программные комплексы УПДК-МК авто и стабилотренажёр Д-01.

Аппаратно-программный комплекс тестирования и развития УПДК-МК авто обеспечивает тестирование следующих ПВК водителя.

1) Оценка психофизиологических характеристик: готовность к психофизиологическим исследованиям; восприятие пространственных отношений и времени; глазомер; устойчивость, переключаемость и распределение внимания; память; психомоторика; эмоциональная устойчивость; динамика работоспособности; скорость формирования психомоторных навыков; оценка моторной согласованности действий рук.

2) Оценка свойств и качеств личности водителя, которые позволяют ему безопасно управлять транспортным средством: нервно-психическая устойчивость; свойства темперамента; склонность к риску; конфликтность; монотоноустойчивость.

Программно-аппаратный стабилометрический комплекс стабилотренажёр Д-01 основан на использовании стабилометрической платформы балансировочного типа с биологической обратной связью по отклонению опорной поверхности от горизонтального положения. Платформа характеризуется наличием ряда устойчивых положений и позволяет оценивать и тренировать способность человека воспроизводить движениями центра тяжести тела заданные траектории.

Оценка уровня психофизиологической пригодности водителей сводится к определению количества уязвимых (ниже нормы) профессионально-важных качеств (ПВК). По результатам соответствия ПВК норме присваивается вид допуска (профпригодности): Д1, Д2, Д3 и недопуск.

Результаты прохождения комплекса тестов представлены на рисунке 4. Испытуемые разбиты на три группы: 1) кандидаты в водители; 2) водители со стажем до 4 лет; 3) водители со стажем 4 и более лет.

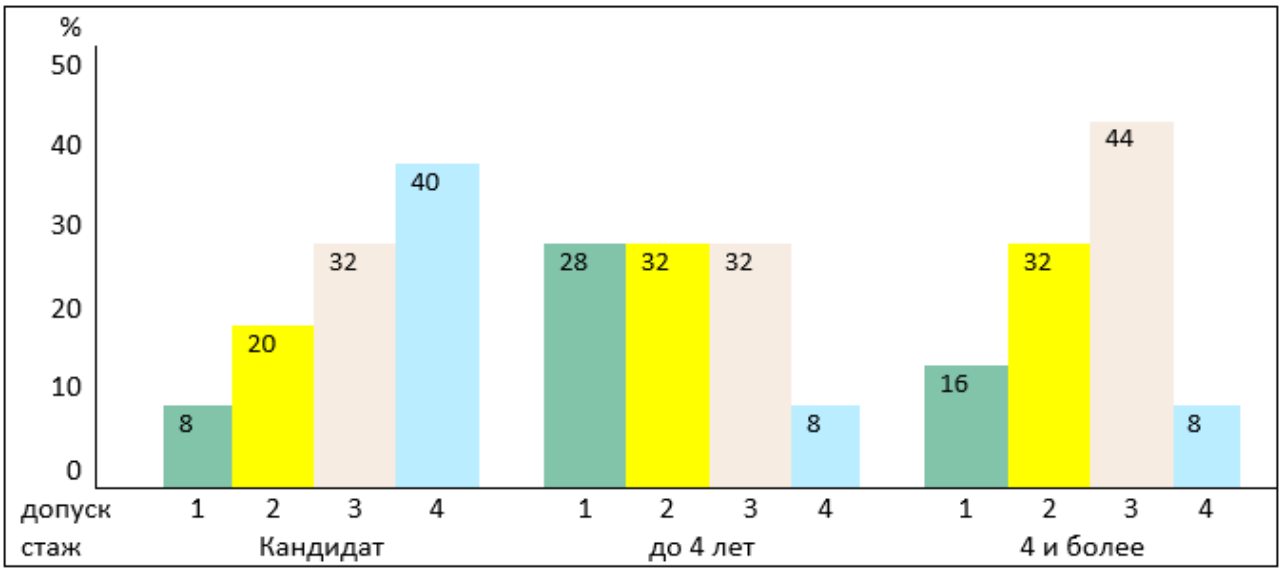


Рис. 4. Результаты прохожденияучастниками эксперимента комплекса тестов, где 1-Д1, 2-Д2, 3-Д3, 4-Недопуск

Процент недопуска снижается по мере возрастания стажа вождения. При этом 3-я группа (стаж вождения более 4 лет) показала, что по мере приобретения опыта происходит снижение самоконтроля за счет проявлений самоуверенности и некоторого пренебрежения к четности выполнения заданий.

Результаты теста на эмоциональную устойчивость показали следующее. Из всех участников эксперимента наиболее эмоционально устойчивыми являются 1-ой и 3-ей группы, т.е. участники, не имеющие опыта вождения (не знакомые с опытом принятия экстренных решений при аварийных ситуациях на дорогах) и водители, имеющие опыт вождения более 4 лет.

В таблице 2 представлен результат обработки результатов оценки уровня психофизиологических качеств и определения соотношения взаимного влияния некоторых психофизиологических качеств между собой.

Таблица 2. Коэффициент корреляции показателей эмоциональной устойчивости и сложной двигательной реакции

СДР-М / ЭУ	Среднее время реагирования в задании №1	Время выбора	Среднее время реагирования в задании №2	Разница средне-арифмети-ческих времен реагирования
Количество ошибок без помехи	-0,40	-0,55	-0,70	-0,56
Среднеарифметическое время реагирования без помехи	0,42	-0,02	0,26	-0,03
Количество пропусков с помехой	0,42	-0,04	0,25	-0,04
Среднеарифметическое время реагирования с помехой	0,36	-0,10	0,16	-0,10

Выявлена обратно пропорциональная корреляционная зависимость между временем реагирования при сложных двигательных зрительно-моторных действиях и показателем эмоциональной устойчивости (количество ошибок без помехи), r = -0,70. Из этого можно сделать вывод, что чем меньше времени тратит водитель при принятии решения в процессе вождения, тем больше вероятность совершения ошибок на этапе принятия решения и при выполнении управляющих действий в нестандартных ситуациях и в условиях действия отвлекающих факторов.

Детальный анализ параметров психофизиологических качеств показывает, что при оценке уровня восприятия скорости и расстояния, количество точных попаданий пропорционально времени реагирования при оценке эмоциональной устойчивости (r = 0,5). Это означает, что чем выше эмоциональная устойчивость, тем более верно водитель транспортных средств способен оценивать скорость и дистанцию во время движения.

Литература

[1]. Дементиенко В.В. и др. Гипотеза о природе электродермальных реакций / Физиология человека. – 2000. – № 2, том 26. – С. 124-131.

[2]. Гедранович Ю. А. Обзор и сравнительный анализ методов и систем для развития навыков релаксации. / Ю. А. Гедранович, В.В. Савченко, К.Д. Яшин, Н.В. Щербина // Журнал «Человеческий фактор: проблемы психологии и эргономики», 2016, № 1 (77), С. 62 – 69. URL: http://elibrary.ru – 22.03.2017.

[3]. Гедранович Ю. А. Обзор и сравнительный анализ методов и систем для развития навыков релаксации. / Ю.А. Гедранович, В.В. Савченко, К.Д. Яшин, Н.В. Щербина // Журнал «Человеческий фактор: проблемы психологии и эргономики», 2016, № 2 (78), С. 44 – 50. URL: http://elibrary.ru – 22.03.2017.

[4]. Применение модуля «Биоадаптивная система NeuroDog» как одного из методов психологической коррекции http://www.neurocom.ru – 6.03.2014.

[5]. Дятлов М.Н. и др. Профессиональная надежность водителя автомобильного транспорта // Молодой ученый. – 2013. – №10. –С. 134-138.

ЛАЗЕРНАЯ БЕЗОПАСНОСТЬ – МЕДИЦИНСКИЕ АСПЕКТЫ

***И.Г. Ляндрес*[1]**
главный специалист по лазерным медицинским технологиям научно-производственного унитарного предприятия «Научно-технический центр «ЛЭМТ» БелОМО», доктор медицинских наук, профессор

***А.П. Шкадаревич*[1]**
директор научно-производственного унитарного предприятия «Научно-технический центр «ЛЭМТ» БелОМО», академик НАН Беларуси, доктор физико-математических наук, профессор

***О.Н. Мартинович*[2]**
инженер-программист, ЧУП «Оптиксофт»

***И.А. Какшинский*[1]**
инженер-технолог 2 категории научно-производственного унитарного предприятия «Научно-технический центр «ЛЭМТ» БелОМО».

[1]*Унитарное предприятие «НТЦ «ЛЭМТ» БелОМО», Республика Беларусь*
[2]*ЧУП «Оптиксофт», Республика Беларусь*

Abstract. Purpose of the work: to ground the possibility of adverse effect of laser radiation on unprotected eye depending on power, wavelength, radiation mode and optical light amplification due to focusing of a crystalline lens. The most dangerous is laser radiation of IR - optical spectral range. Protection degree of eyes by goggle's light filters is defined by their optical density: optical power attenuation shouldn't exceed 1 mW for visible range, and it has to be maximum (OD-3, OD-6) for IR range.

Американский институт стандартов(ANSI) разделил лазеры на четыре класса по степени их опасности. Эта классификация легла в основу национальных классификаций.

В Республике Беларусь действуют санитарные правила и нормы 2.2.4.1.13-2-2006 «Лазерное излучение и гигиенические требования при эксплуатации лазерных изделий», в которых применяется указанная выше классификация.

Она базируется на учете выходной мощности (энергии) лазерного излучения (ЛИ) и предельно-допустимых уровнях (ПДУ) лазерного облучения глаз и кожи.

Различают лазерные изделия закрытого и открытого типа. Закрытые лазерные изделия исключают прямое воздействие на человека.

У открытых лазерных изделий ЛИ выходит в рабочую среду, претерпевая отражение, поглощение, обратное рассеивание и другие оптические эффекты в зависимости от оптических характеристик среды.

Особенно опасно ЛИ для глаз при прямом воздействии.

Различают однократное действие ЛИ и хроническое. Однократное воздействие, как правило, случайное и кратковременное, хроническое – связано с профессиональной деятельностью и представляет опасность при несоблюдении мер защиты. Фотофизические и фотобиологические эффекты действия ЛИ определятся оптическими характеристиками биоткани, параметрами и режимами работы лазерных изделий.

Возможны термическое, фотохимическое, фотоакустическое воздействие на органы зре-

ния, а также влияние на организм электромагнитного излучения, ионизации воздуха, вдыхания продуктов горения и испарения материалов, на которые падает лазерный луч, фотоожоги кожи.

Термическое воздействие на глаз обусловлено с одной стороны свойствами ЛИ, с другой – оптическими характеристиками сред глаза, которые являются прозрачными для видимого и инфракрасного (ИК) диапазонов спектра. Следует учитывать, кроме того, что сетчатка обладает свойствами селективного поглощения в зеленом и, в меньшей степени, красном спектре.

Наименее опасным является ЛИ лазера первого класса с мощностью в пределах 1 мВт. Однако, даже кратковременное прямое воздействие луча такой мощности, если глаз открыт, может вызвать фотоофтальмию. При нормальном освещении зрачок сужен, фотоофтальмия носит временный характер. При низкой освещённости, когда зрачок расширен, может наступить более длительное снижение остроты зрения, особенно, если воздействовал зеленый свет лазера. Излучение лазеров терапевтического диапазона за счет фокусировки хрусталика дает оптическое усиление х105 (рисунок 1).

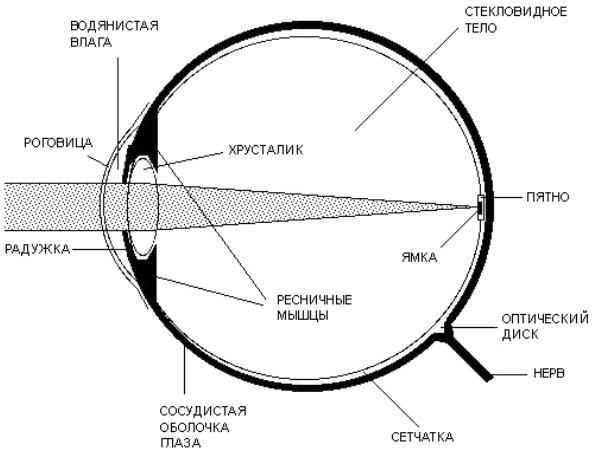


Рис. 1. Потенциальная опасность лазерного излучения незащищенного глаза за счет фокусировки хрусталиком [4]

Это означает, что при плотности мощности 1 мВт/см^2 на сетчатке формируется световое пятно с плотностью мощности до 100 мВт/см^2, если луч коллимированный.

Лазеры терапевтического диапазона с мощность до 500 мВт работают во многих лечебных учреждениях и имеют открытый луч. Расчеты показывают, что плотность мощности ЛИ таких лазеров достаточна, чтобы вызвать термическое повреждение сетчатки, а также фотохимические реакции. Например, луч в красном диапазоне спектра мощностью 60 мВт, сфокусированный на сетчатке в виде светового пятна диаметром до 50 мкм, формирует плотность мощности 3 Вт/см^2, а термодеструкция сетчатки может наступить даже при меньшей мощности и приобрести необратимый характер.

Наряду с мощностью большое значение имеет проникающая способность в ткани ЛИ различных длин волн. Эти свойства связано с особенностями взаимодействия ЛИ и основных хромофоров биотканей человека. Они же определяют селективность лазерного воздействия.

Основными хромофорами биотканей человека являются вода, гемоглобин (оксигемоглобин), пигменты. Все три хромофора в тканях глаза присутствуют: вода содержится в жидких средах глаза, роговице и хрусталике, гемоглобин и пигменты – в сетчатке.

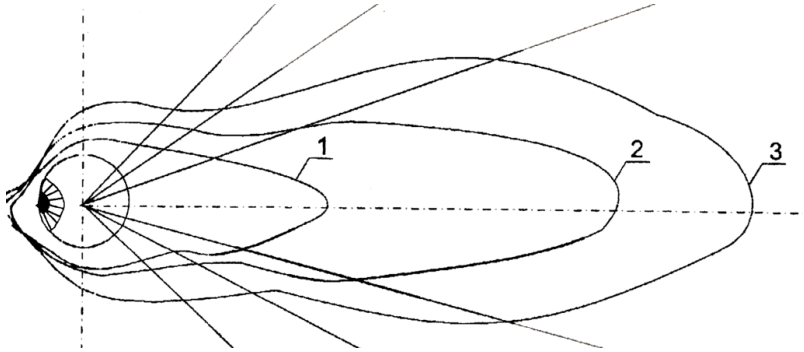


Рис. 2. Оптические характеристики энуклеированного глаза при воздействии лазерного излучения с длиной волны 630 нм (1), 1150 нм (2), 890 нм (3) [2]

Из рисунка 2 следует, что наибольшей проникающей способностью обладает излучение с длиной волны 890 нм («окно биологической прозрачности» ткани). Излучение с длиной волны 1150 нм является водоспецифичным и частично поглощается жидкими средами глаза. В зеленом диапазоне спектра опасность повреждения сетчатки существенно увеличивается (в связи с селективным поглощением пигментом сетчатки) при значительно меньшей мощности, чем в красном диапазоне.

Особенно опасен импульсный режим ЛИ, в частности, Q switched импульсы.

Наиболее неблагоприятными для глаза являются длины волн от 380 до 1400 нм, так как они слабо поглощаются водосодержащими тканями глаза и фокусируются на поверхности сетчатки, создавая большую плотность мощности.

Излучение гольмиевых лазеров (λ 2094 нм) частично поглощается жидкостью камер глаза, но риск повреждения сетчатки остается, особенно при Q – модулированном импульсном режиме.

Эрбиевый лазер (λ 2940 нм) генерирует излучение с высоким коэффициентом поглощения водой и представляет опасность для передних камер глаза. Это же относится к CO_2 (λ 10600 нм) лазеру.

Излучение лазеров первого и второго классов видимого диапазона (400-700 нм) мощностью менее 1 мВт при экспозиции 0.25 секунды (время, за которое человек успевает закрыть глаза или отвернуться) не повреждает сетчатку.

Ультрафиолетовое излучение, в зависимости от дозы и частотных характеристик, приводит к повреждению роговицы.

При прямом или отраженном излучении лазеров 3А, 3В классов или диффузном отражении излучения лазеров 4 класса повышенной мощности повреждения могут произойти, прежде чем человек рефлекторно закроет глаза.

Импульсный режим излучения негативно действует на глаз при длительности импульсов менее 1 мс. При этом возникает нарушение кровообращения в сосудах сетчатки, а также их повреждение с кровоизлиянием в стекловидное тело.

При импульсном режиме имеет место фотоакустический эффект, который также может вызвать повреждение сетчатки.

Фотохимическая реакция со стороны сетчатки может иметь место при длительном воздействии ЛИ на незащищённый глаз в фиолетовом диапазоне (400-470 нм).

Защита глаз от лазерного излучения требует ношения специальных очков со светофильтрами. Степень защиты светофильтрами определяется оптической плотностью - optical density (OD), которая показывает, во сколько раз происходит ослабление света. OD-1 означает ослабление в 10 раз, OD-3 в 1000 раз, OD-6 – в 1000000 раз. Для видимого диапазона OD должна быть такой, чтобы мощность излучения падающего на глаз, не превышала мощности излучения второго класса лазера (≈ 1 мВт). Это связанно с необходимостью наблюдать луч в качестве

пилотного для наведения на объект.

Для ИК-диапазона степень ослабления должна быть максимальной, так как излучение этого диапазона обладает большой проникающей способностью.

В санитарных правилах и нормах [3], указаны следующие марки светофильтров: СЗС-22 – диапазон защиты 630-680 нм; 680-1200 нм;1200-1400 нм; OD – соответственно 3, 6, 3 цвет стекла – голубой (рисунок 3). Для защиты от сине-зеленого спектра используются светофильтры ОС-23, цвет стекла – оранжевый. От излучения CO_2 и эрбиевых лазеров защищают очки с обычными стеклами.

Для лазеров, излучающих в ИК диапазоне, использующих красный пилотный луч, необходимы защитные очки, экранирующие ИК диапазон и, одновременно, оставляющие видимым красный свет, что важно при проведении оперативных вмешательств.

На светофильтрах защитных очков необходимо наличие маркировки, указывающей на диапазон длин волн, от которых защищают светофильтры.

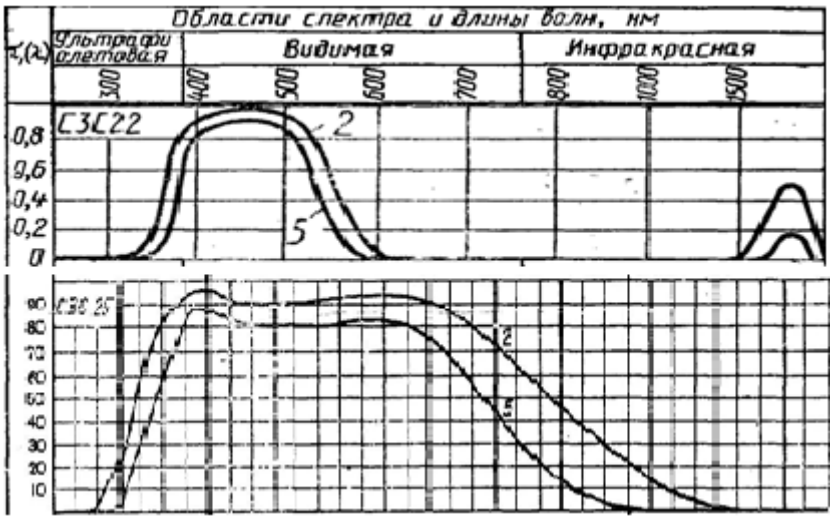


Рис. 3. Кривые пропускания длин волн лазерного излучения светофильтрами защитных очков СЗС-22 и СЗС-25

Показано, что при хроническом воздействии лазерного излучения на работающих имеет место снижение артериального давления. Поэтому работникам с низким артериальным давлением не рекомендуется обслуживать лазерные установки.

Гипотензивный эффект может быть вызван также электромагнитным полем, создаваемым мощным лазером.

Импульсный режим работы мощных лазерных систем может спровоцировать нарушение ритма сердечных сокращений у больных с ишемической болезнью сердца, осложнённой периодическими нарушениями сердечного ритма. Это же касается пациентов с установленными кардиостимуляторами. Допуск таких работников к работе на лазерных установках противопоказан.

При несоблюдении технологии работы с лазером возможны термические повреждения кожи, если мощность ЛИ превышает 1 Вт. У людей с темным цветом кожи термическое повреждение возможно и при меньшей мощности ЛИ за счет большего поглощения. Термические лазерные ожоги, как правило, возникают редко и связаны с нарушением техники безопасности (рисунок 4)

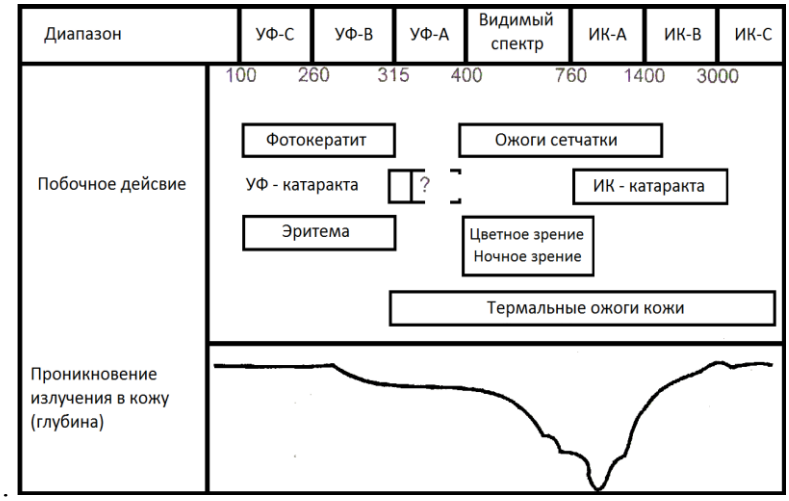


Рис.4. Фотобиологические спектральные диапазоны, разработанные Международной комиссией по люминесценции, и их воздействие на структуры глаза и кожу [1]

ЛИ в УФ диапазоне при хроническом воздействии на незащищённую кожу может вызвать её пигментацию вплоть до появления новообразований.

Заключение. Лазерное излучение неблагоприятно воздействует на незащищённый глаз благодаря оптическому усилению за счет фокусировки хрусталиком.В зависимости от длины волны, селективности, импульсного характера лазерного излучения фотодеструктивное воздействие на элементы глаза может быть различным.Использование защитных очков со специальными светофильтрами обеспечивает надежную защиту глаз от лазерного излучения. Степень защиты светофильтрами определяется их оптической плотностью: для видимого диапазона ослабление излучения должно соответствовать мощности 1мВт, для ИК диапазона оно должно быть максимальным(OD-3;OD-6).

Литература

[1]. Ляндрес И.Г. Низкоинтенсивные лазеры в клинической практике / Ляндрес И.Г.. – Минск, 1998г. – 227 с.

[2]. Евстигнеев А.Р., Полонский А.К. Применение полупроводниковых лазеров и светодиодов в биомедицине и медицинском приборостроении / Евстигнеев А.Р., Полонский А.К.. – Калуга,1989г. – ,155с.

[3]. Санитарные правила и нормы 2.2.4.1.13-2-2005 «Лазерное излучение и гигиенические требования при эксплуатации лазерных изделий». – Минск, 2006г.

[4]. Безопасность при работе с лазерами и что будет, если её не соблюдать [Электронный ресурс]. – Режим доступа: http://lasers.org.ru/2008/06/20/ – Дата доступа – 3.21.2016.

ОПЫТ ПРЕПОДАВАНИЯ ДИСЦИПЛИНЫ АЛГОРИТМЫ МАШИННОГО ОБУЧЕНИЯ В ВУЗЕ

М.А. Амелин
Асистент кафедры инженерной психологии и эргономики БГУИР, магистр экономических наук

Белорусский государственный университет информатики и радиоэлектроники, Республика Беларусь
E-mail: michael.amelin@gmail.com

Abstract. recent years, the use of machine learning methods has significantly increased and therefore research in this field is becoming more and more important. Competitiveness and success of organizations in the global economy will depend on the ability to use the methods of machine learning. Using of these methods will help to optimize those business processes in which it is necessary to analyze big data and automate administrative decisions.

В последние годы, использование методов машинного обучения значительно возросло и поэтому исследования в этой области становятся всё более важными. От умений качественно применять методы машинного обучения будут зависеть конкурентоспособность и успешность предприятий в глобальной экономике. С помощью методов машинного обучения предприятия получают возможность оптимизировать те свои процессы, в которых необходимо анализировать большие данные и автоматизировать управленческие решения.

В ходе подготовки к занятиям в 2016 / 2017 учебном году по дисциплине «Алгоритмы машинного обучения» автором статьи было пройдено обучение, сданы экзамены и получены сертификаты по следующим курсам, связанным с контекстом больших данных и алгоритмов машинного обучения:

1 Big Data Fundamentals, Armonk, New York, IBM (BDU), 2016;
2 Introduction to R-DataCamp, Armonk, New York, IBM (BDU), 2016;
3 Introduction to Machine Learning, Boston, DataCamp, 2016;
4 Intermediate R, Boston, DataCamp, 2016;
5 Intro to Statistics with R: Correlation and Linear Regression, Boston, DataCamp, 2016;
6 Hadoop Fundamentals I, Armonk, New York, IBM (BDU), 2016;
7 Big Data with SAP HANA Vora, Walldorf, SAP, 2016;
8 Enterprise Machine Learning in a Nutshell, Walldorf, SAP, 2016;
9 Imagine IoT, Walldorf, SAP, 2016;
10 Machine Learning for Data Science, San Francisco, Udemy, 2016;
11 Hands-on Industry 4.0, Potsdam, Hasso Plattner Institute for Software Systems Engineering (MOOC.House), 2016.

На рисунках 1–6 представлены некоторые из сертификатов, полученных во 2-й половине 2016 года.

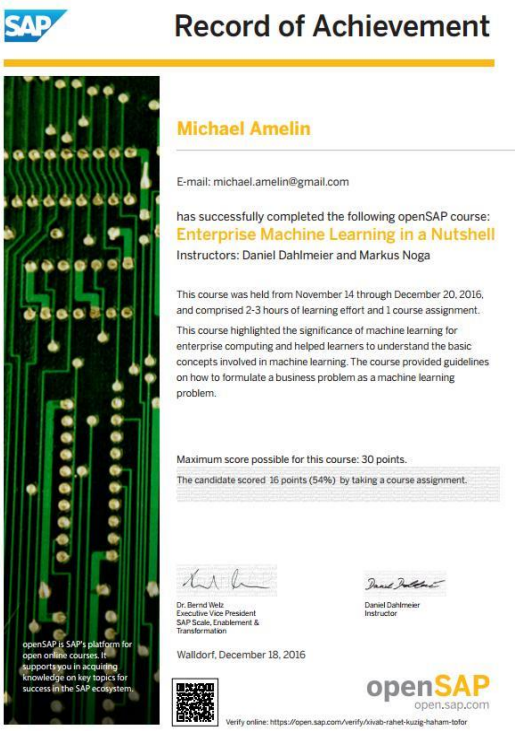


Рис. 1. Сертификат Enterprise Machine Learning in a Nutshell

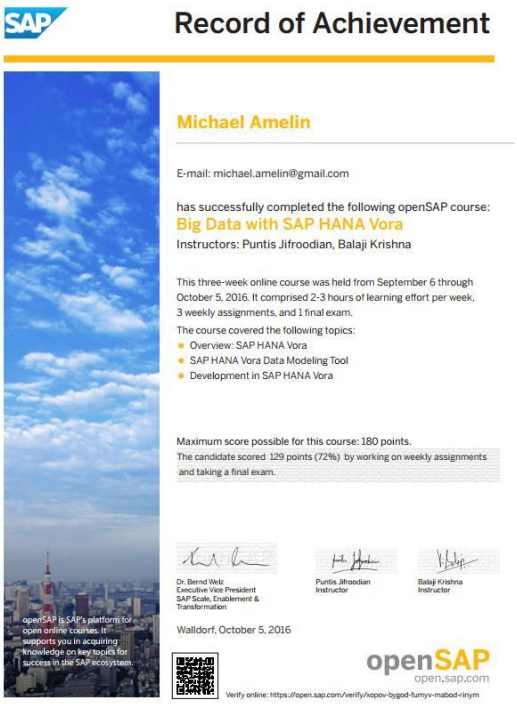


Рис. 2. Сертификат Big Data with SAP HANA Vora

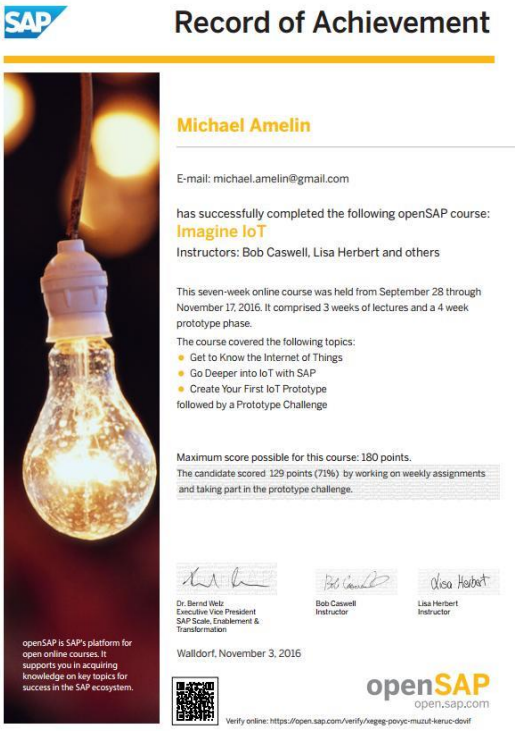


Рис. 3. Сертификат Imagine IoT

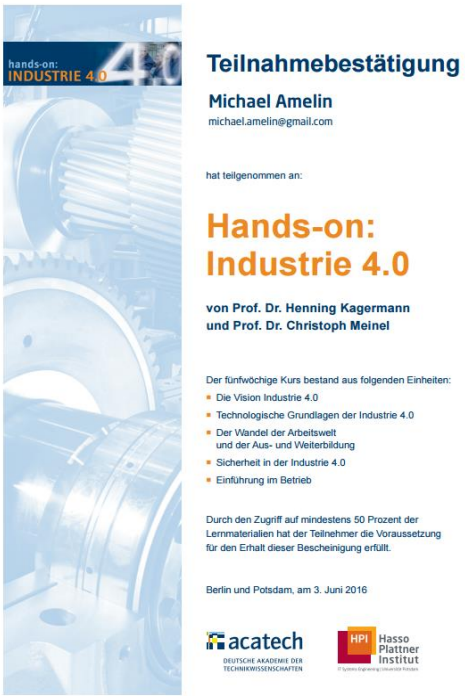


Рис. 4. Сертификат Hands-on: Industire 4.0

Рис. 5. Сертификат Introduction to Machine Learning

Рис. 6. Сертификат Machine Learning for Data Science

Проводя экскурс в историю происхождения термина большие данные нужно упомянуть, что под ним подразумевается набор данных, который является настолько большим или сложным, что традиционные приложения обработки данных недостаточны для усвоения их. Также проблемы включают в себя анализ, поиск, совместное использование, хранение, передачу, визуализацию, обработку запросов, обновление и конфиденциальность информации. Термин «большие данные» часто относится просто к использованию прогностического анализа, аналитике поведения пользователей, и некоторым другим передовым методам анализа информации [1].

Организации могут повысить показатели своей эффективности и получить конкурентные преимущества при помощи надлежащей бизнес-аналитики. Существует четкая положительная корреляция между бизнес-анализом и предпринимательским успехом [2]. Увеличивающееся разнообразие типов данных, их источников, а также объема и скорости даёт организациям возможность для получения большего количества информации и последующего принятия обоснованных бизнес-решений.

Аналитика больших данных дополняет традиционные статические отчеты и помогает получить конкурентное преимущество организациям за счет верного предсказания бизнес-ситуации, оптимизации и улучшения адаптивности бизнес-модели [3]. Тем не менее, управление качеством данных становится всё более сложным, так как их разнообразие и количество источников постоянно увеличивается.

Большие данные имеют тенденцию постоянно расти в своём объеме, скорости и диапазоне типов и источников. При этом низкое качество данных является растущей проблемой. Наличие ошибок и несоответствий в источниках бизнес-данных и неправильная частота их сбора зачастую всё ещё не учитываются при проведении их аналитики.

Целью традиционной бизнес-аналитики является создание экономической ценности, своевременное нахождение существенных изменений в бизнес-процессах и принятие верных решений. Задачей же аналитики больших данных является извлечение полезной информации из массивных хранилищ данных. Такая информация может быть извлечена с помощью качественного сопоставления, в котором явление изучается путем наблюдения и выводов о переменных, которые измеряют сам феномен.

Описывая участие студентов в подготовке к сдаче экзамена по дисциплине «Введение в анализ данных с использованием технологии R» (Big Data University (BDU) от IBM), отметим, что они прошли онлайн-практику на электронной образовательной платформе DataCamp.

Рис. 7. Иллюстрации сертификатов студентов, прошедших онлайн-обучение на образовательной платформе DataCamp (BDU partner), Boston, USA

Также кроме прохождения курсов по «Основам BIGDATA», «Введение в анализ данных с использованием технологии R» (BDU) и «Введения в язык программирования R» (DataCamp) некоторые студенты смогли пройти следующие образовательные курсы на платформе Big Data University от IBM.

Рис. 8. Примеры сертификатов студентов, прошедших онлайн-обучение на образовательной платформе BDU

Далее рассмотрим некоторые вопросы, которые были предложены студентам для изучения разделов BIGDATA (в контексте алгоритмов машинного обучения) на лекционных, практических и лабораторных занятиях в БГУИР:

– Что такое Машинное обучение?

– Интеллектуальные приложения, работающие при помощи алгоритмов машинного обучения.

– Переход от бизнес-проблемы к задаче машинного обучения.

– Машинное обучение в контексте корпоративных вычислительных систем.

– Методы классификации в машинном обучении.

– Логистическая и линейная регрессия в машинном обучении.

– Примеры ИТ-приложений с учётом использования алгоритмов машинного обучения.

– Практические сферы применения алгоритмов машинного обучения.

Рис. 9. Слайды лекций «Алгоритмы машинного обучения»

В ходе следущих аудиторных занятий студентам также были предложены варианты расширения их знаний в области экосреды больших данных:

Рис. 10. Слайды лекций «Введение в большие данные»

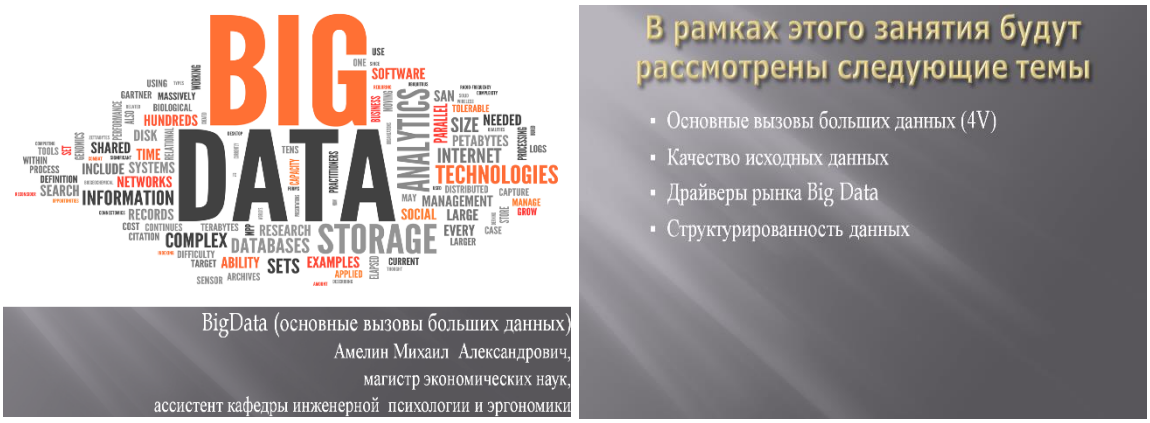


Рис. 11. Слайды лекций «Основные вызовы больших данных»

Рис. 12. Слайды лекций «Системные технологии для больших данных»

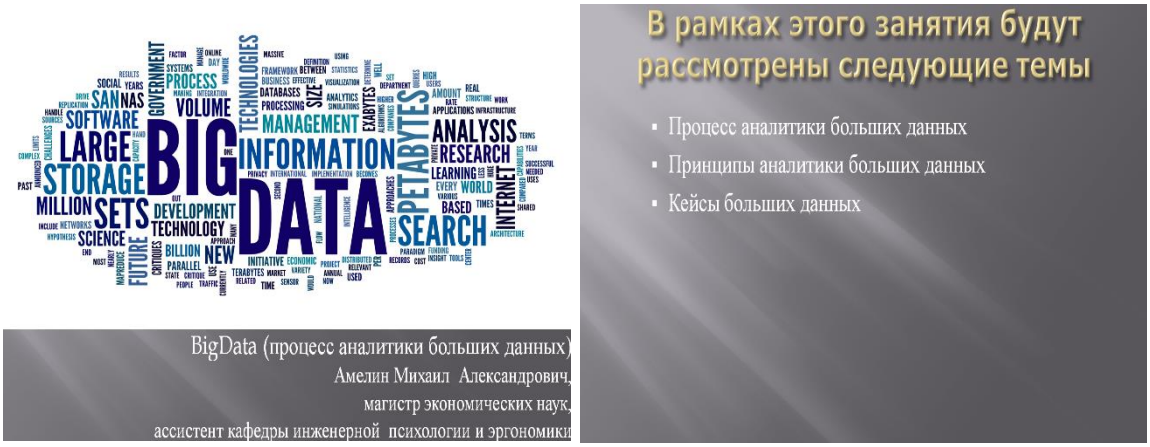


Рис. 13. Слайды лекций «Процесс аналитики больших данных»

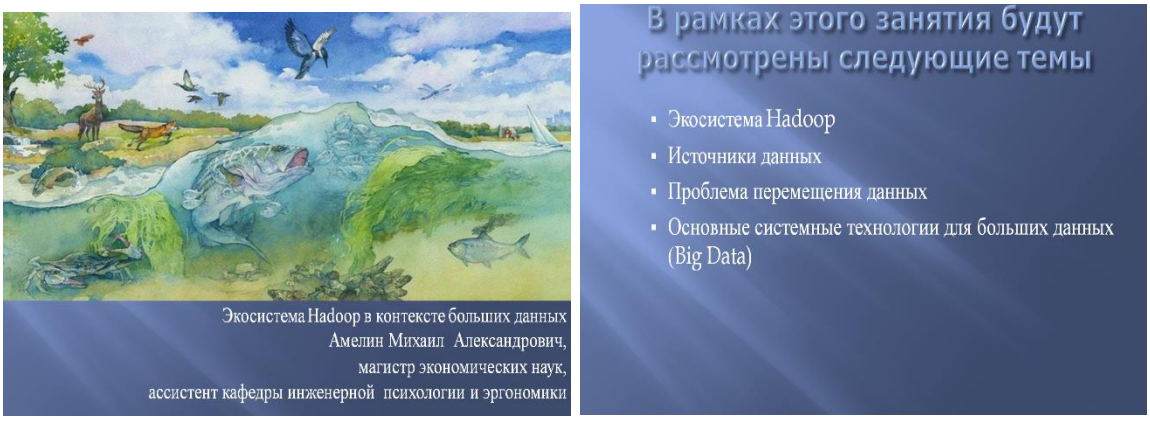


Рис. 14. Слайды лекций «Экосистема Hadoop в контексте больших данных»

Рис. 15. Слайды лекций «Методы анализа больших данных»

Во втором семестре 2016–2017 учебного года, в апреле и мае, две группы студентов факультета компьютерного проектирования (БГУИР) пройдут обучение в BIG DATA UNIVERSITY по программе AmbassadorBDU от IBM, DataCamp и др. Ожидаются группы 410901 (3 курс, специальность Инженерно-психологическое обеспечение информационных технологий, инженеры-системотехники, 32 человека) и 410101 (3 курс, специальность Информационные системы и технологии (в обеспечении промышленной безопасности), инженеры-системотехники, 26 человек). Обе группы – очное, 4-летнее обучение.

Приведём несколько примеров лабораторных занятий для студентов БГУИР по языку программирования R, в контексте подговки к решению задач машинного обучения:

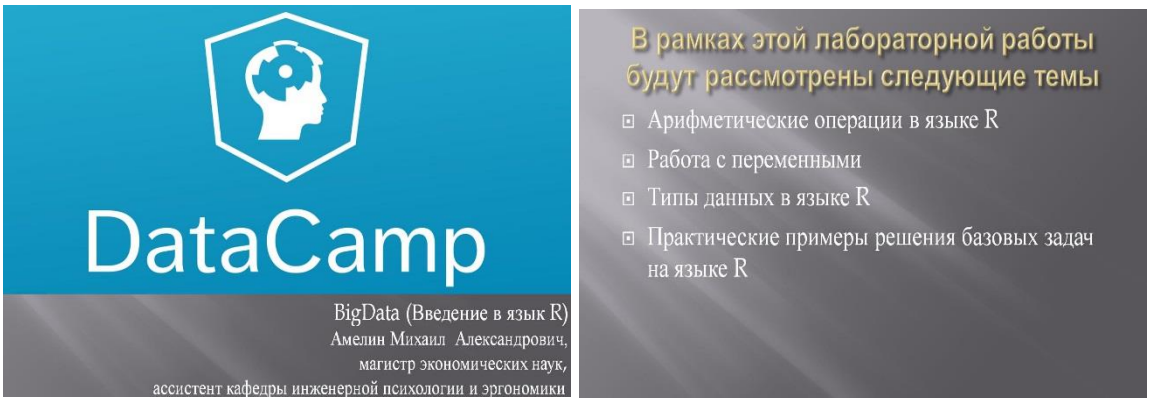


Рис. 16. Слайды лабораторных работ «Введение в язык R»

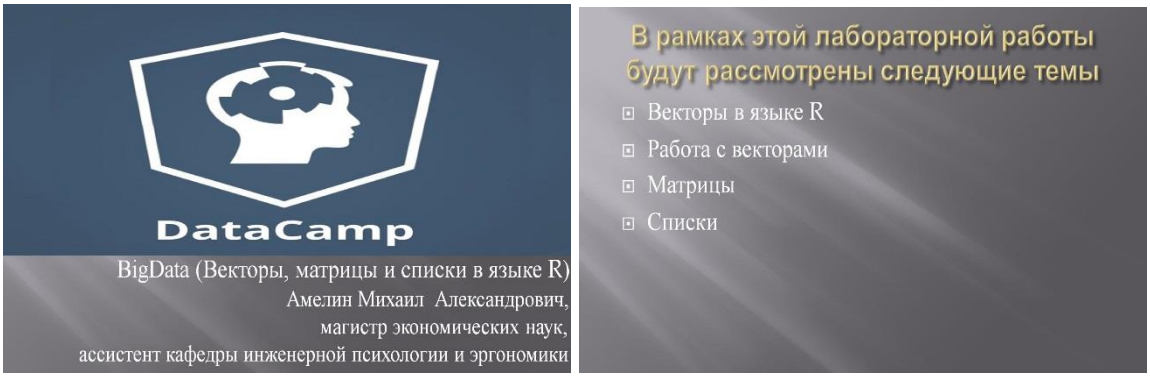


Рис. 17 Слайды лабораторных работ «Векторы, матрицы и списки в языке R»

В заключение добавим, что дальнейшая подготовка студентов на практических и лабораторных занятиях проводилась с учётом освоения следующих тем:

- Условные выражения и управление потоком данных;
- Логические операторы в языке R;
- Условные операторы;
- Цикл While;
- Цикл For;
- Введение в функции.

В результате курса «Алгоритмы машинного обучения» студенты готовы практически подходить к решению таких задач как: создание базовой модели прогнозирования, работе с матрицей неточностей, кластеризации видов, классификации и фильтрации спама в почте, моделирование будущих просмотров учётных записей в социальной сети LinkedIn и др.

Литература

[1]. Амелин, М.А. Аналитика больших данных: вызовы и решения / М.А. Амелин, А.В. Артухевич // Topical areas of fundamental and applied research X: материалы X Международной науч.-практ. конф., NorthCharleston, 7–8 ноября 2016 г. : в 3 ч. / CreateSpace. – North Charleston, USA, 2016. Vol. 2. – С. 165–167.

[2]. Maheshwari, A. Data Analytics Made Accessible / A. Maheshwari. – Seattle: Amazon Digital Services, 2014. – 156 p.

[3]. Ankam, V. Big Data Analytics / V. Ankam. – Birmingham: Packt Publishing, 2016. – 326 p.

КОМПЬЮТЕРНЫЕ СИСТЕМЫ КАК ОБЪЕКТ ИНЖЕНЕРНО-ПСИХОЛОГИЧЕСКОГО ИССЛЕДОВАНИЯ

Е.Д. Азаренко
Аспирант кафедры инженерной психологии и эргономики БГУИР

Т.Ю. Шлыкова
Доцент кафедры инженерной психологии и эргономики БГУИР, кандидат психологических наук, доцент

Белорусский государственный университет информатики и радиоэлектроники, Республика Беларусь
E-mail: aifabregaz@gmail.com, ty_shlykova@mail.ru

Abstract. The article describes computer systems from the point of view of monitoring process. Nowadays, amount of data on servers is growing and service providers have to ensure service level agreements (SLA) for the customers. At the moment number of monitoring engineers is growing much slower than number of servers and storages on the server side. That fact forces to search the way of monitoring process optimization. Engineers that perform monitoring should track only information that really makes sense. Computer systems are analyzed to find their monitoring parameters then automated monitoring systems and ways of optimization were provided.

Современные компьютерные системы с каждым днем должны выдерживать все большие нагрузки, связанные с обработкой и хранением данных. Растущие объемы данных требуют обеспечения надежного хранилища, поэтому на современных предприятиях происходит модернизация, которая включает увеличение числа вычислительных ресурсов и памяти. Помимо этого, конкурентная среда в индустрии информационных технологий требует от компаний предоставления пользователям круглосуточного стабильного доступа к предлагаемым услугам.

При предоставлении услуг облачного сервиса или хранилища, компания-провайдер гарантирует определенный уровень качества услуг (Service Level Agreement (SLA)). Он включает в себя требования по быстрому отклику службы технической поддержки провайдера и жестко заданный временной уровень доступности сервиса в процентах (напр., 99,5% для Amazon Web Services) [1].

На предприятиях для обеспечения постоянной доступности компьютерных систем осуществляется их мониторинг, которым занимается команда квалифицированных инженеров. Инженеры, занимающиеся мониторингом систем, производят контроль вычислительных систем и восстановление их работоспособности, для чего требуются не только сосредоточенность, но и аналитические навыки, которые способствуют предотвращению внештатных ситуаций. На рынке труда во всем мире наблюдается нехватка специалистов, имеющих необходимую квалификацию.

Следовательно, с одной стороны мы наблюдаем увеличение количества вычислительных мощностей на предприятиях, с другой – высокий неудовлетворенный спрос на инженеров по мониторингу компьютерных систем на рынке труда. В таких условиях в большинстве компаний при увеличении количества серверов и показателей мониторинга систем не происходит пропорционального увеличения количества инженеров, что ведет к увеличению нагрузки на штат и скорейшему эмоциональному выгоранию сотрудников [2].

Данные обстоятельства вынуждают искать пути оптимизации инженерно-психологического мониторинга с целью уменьшения нагрузки на оператора. И чтобы сократить поток отслеживаемых инженером данных, обратимся к объекту инженерно-психологического мониторинга – компьютерным системам.

Компьютерные системы обладают аппаратной и программной частью, каждая из которых, в свою очередь, имеет свои показатели, определяющие состояние и производительность системы.

Аппаратные параметры компьютерных систем позволяют следить за состоянием физических частей объекта мониторинга. Аппаратное обеспечение сервера во многом определяет производительность сервиса. При инженерно-психологическом мониторинге необходимо контролировать такие параметры, как загрузку процессора, количество занятой и свободной оперативной памяти, количество свободного дискового пространства, уровень изношенности дисковых накопителей, уровень загруженности сетевого канала.

Вышеперечисленные параметры необязательно отслеживать в виде графиков, так как при правильно подобранной конфигурации сервера он должен иметь запас производительности.

Программные параметры компьютерных систем включают характеристики работы приложений, развернутых на компьютерной системе. Рассмотрим в качестве приложения, развернутого на сервере, веб-сервис, который принимает HTTP-запросы от клиентов. В данном случае важно контролировать следующие параметры приложения: среднее время обработки запроса, количество запросов, обрабатываемое за дельта-промежуток времени (напр., минуту), количество запросов, находящихся в очереди на обработку.

Все эти метрики требуют графического отслеживания. График позволяет иметь несколько подгрупп в рамках одного параметра, чем можно эффективно воспользоваться для разделения типов запросов на быстро выполняющиеся и долгие.

Также информативность графиков можно значительно повысить с помощью использования логарифмической шкалы. Наглядную разницу можно получить, сравнив графики, изображенные на рис. 1 и рис. 2. Логарифмическая шкала очень удобна для отображения очень больших диапазонов значений величин, появление на графике пиковых значений не ухудшает детализацию остального потока информации, как в случае линейной шкалы.

Помимо состояния сервера приложения очень полезно отслеживать характеристики, связанные с пользователями. К ним можно отнести количество уникальных пользователей в сутки, среднее время сессии и другие параметры, определяющие качество сервиса для конечных пользователей.

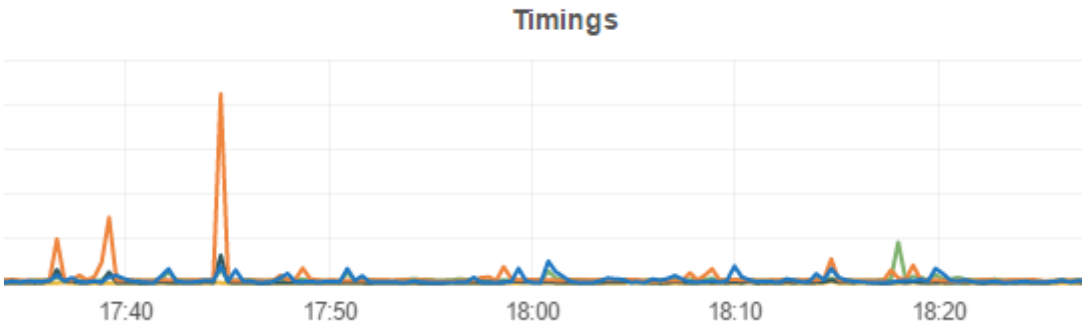


Рис. 1. Использование линейной шкалы для графика, описывающего среднее время обработки запроса.

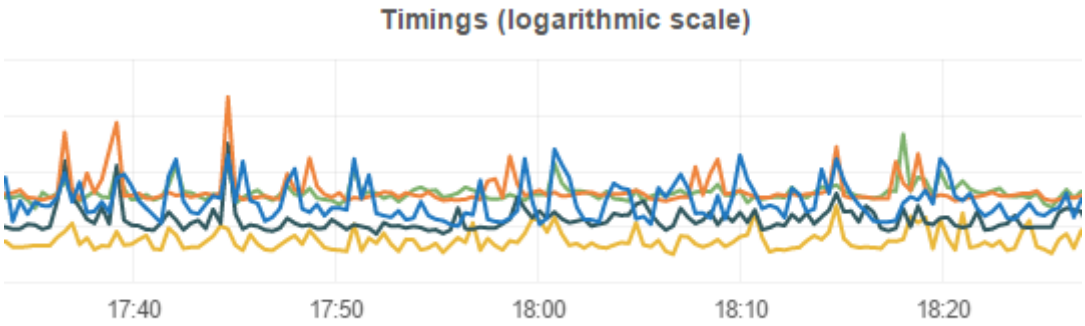


Рис. 2. Использование логарифмической шкалы для графика, описывающего среднее время обработки запроса.

Такой вид инженерно-психологического мониторинга называется мониторингом реальных пользователей (Real User Monitoring (RUM)) и базируется на отслеживании пользовательского трафика [3]. Это может быть сделано посредством анализа сетевого соединения, вставки специальных скриптов в веб-страницы либо установки специальных приложений на пользовательские компьютеры.

В промышленных решениях, обрабатывающих большие потоки информации, для инженерно-психологического мониторинга используются программные продукты, наиболее распространенными из которых являются Nagios, Zabbix, Graphana / Graphite, Solarwinds Server & Application monitor. На данный момент наибольшее распространение получила система Zabbix [4].

Если комплексно рассмотреть список параметров компьютерных систем с точки зрения инженерно-психологического мониторинга, можно предоставить несколько способов оптимизации потока количественных данных. С одной стороны, оптимизация может быть достигнута за счет увеличения информативности графиков состояния компьютерной системы, с другой – за счет отказа от графиков в пользу уведомлений при достижении критических значений тех или иных параметров.

Литература

[1]. Velte A.T. - Cloud Computing: A Practical Approach / Anthony T. Velte - McGraw-Hill, 2010 – 334 с.

[2]. Вайнштейн, Л. А. Психология труда: Курс лекций. / Л. А. Вайнштейн. - Минск: БГУ, 2011 – 219 с.

[3]. Croll A., Power S. Complete Web Monitoring / Alistair Croll, Sean Power - O'Reilly Media, 2009 – 672 с.

[4]. Далле Вакке А. Zabbix. Практическое руководство / А. Далле Вакке - М: ДМК Пресс, 2017. – 356 с.

BIG DATA АНАЛИТИКА И ЕЕ ПРИМЕНЕНИЕ

В.С. Дроздов
Ассистент кафедры инженерной психологии и эргономики, аспирант БГУИР, магистр технических наук

Белорусский государственный университет информатики и радиоэлектроники, Республика Беларусь
E-mail: fxtraid@tut.by

Abstract. Big data provides the opportunity to discover phenomena and laws, which until now was beyond our understanding; linking various aspects of the system to help you better understand their behavior; to assist in the description of complex phenomena and processes; coordinate descriptions of complex systems and enables you to build models that predict their dynamic behavior. These capabilities are deeply relevant to many research fields such as weather forecasting and climate, studies of the brain; determining the state of the global economy; the assessment of performance in agriculture; demographic projections

Big Data – это группа технологий и методов производительной обработки очень больших объемов данных, в том числе неструктурированных, в распределенных информационных системах, обеспечивающих организацию качественно новой полезной информации. Технологии Big Data предоставляют услуги, позволяющие раскрыть потенциал мегамассивов данных за счет выявления скрытых закономерностей и фактов. Под очень большими наборами данных подразумеваются данные объемом от терабайт до сотен петабайт. Рассмотрим важность использования Big Data в различных сферах жизни и производства.

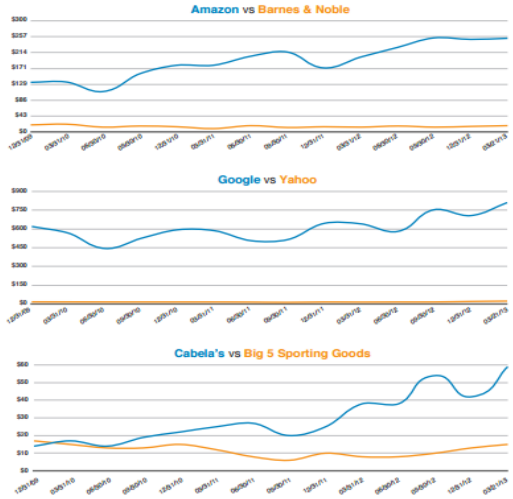
Признаки, характеризующие Большие Данные

Признаки	Пояснение
Объем (volume)	Представляет собой большой объем информации, который трудоемко обрабатывать и хранить традиционными способами
Скорость (velocity)	Данный признак указывает на скорость обработки данных, что в последнее время более востребовано
Многообразие (variety)	Возможность одновременной обработки структурированной и неструктурированной разноформатной информации
Достоверность данных (veracity)	Все большее значение пользователи стали придавать значимость достоверности имеющихся данных
Ценность накопленной информации (value)	Большие данные должны быть полезны компании и приносить особую ценность для нее

Почему Big Data аналитика так важна?

1. Производительность от управления данными

Компании,которые используют Big Data аналитику (синии линии) более продуктивны, чем те компании, которые этого не делают



2. Результаты использования Big Data аналитики

Компании из всех отраслей промышленности используют большие данные для:

- Увеличения выручки
- Сокращения затрат
- Повышения продуктивности

Увеличение доходов (Игорная онлайн компания)

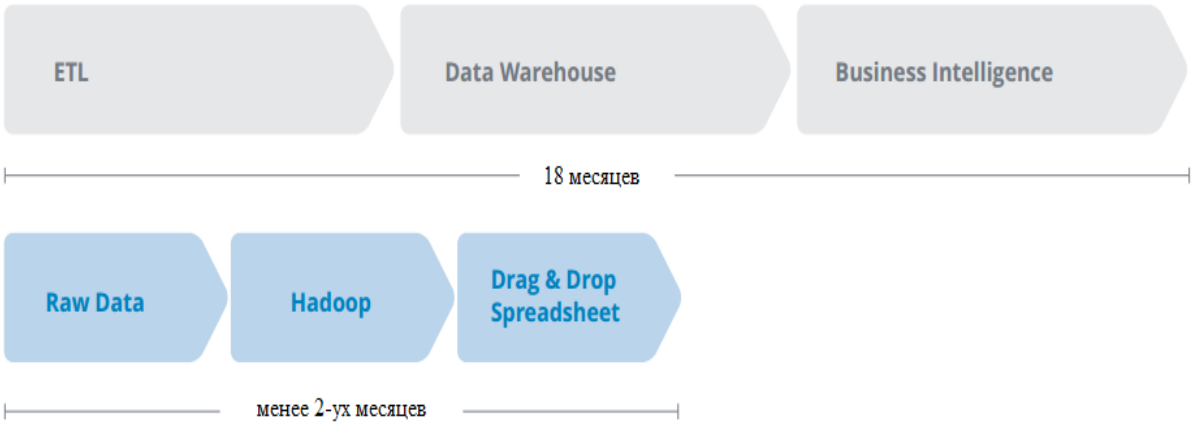
Увеличение числа клиентов(Компания по безопасности в ПО)

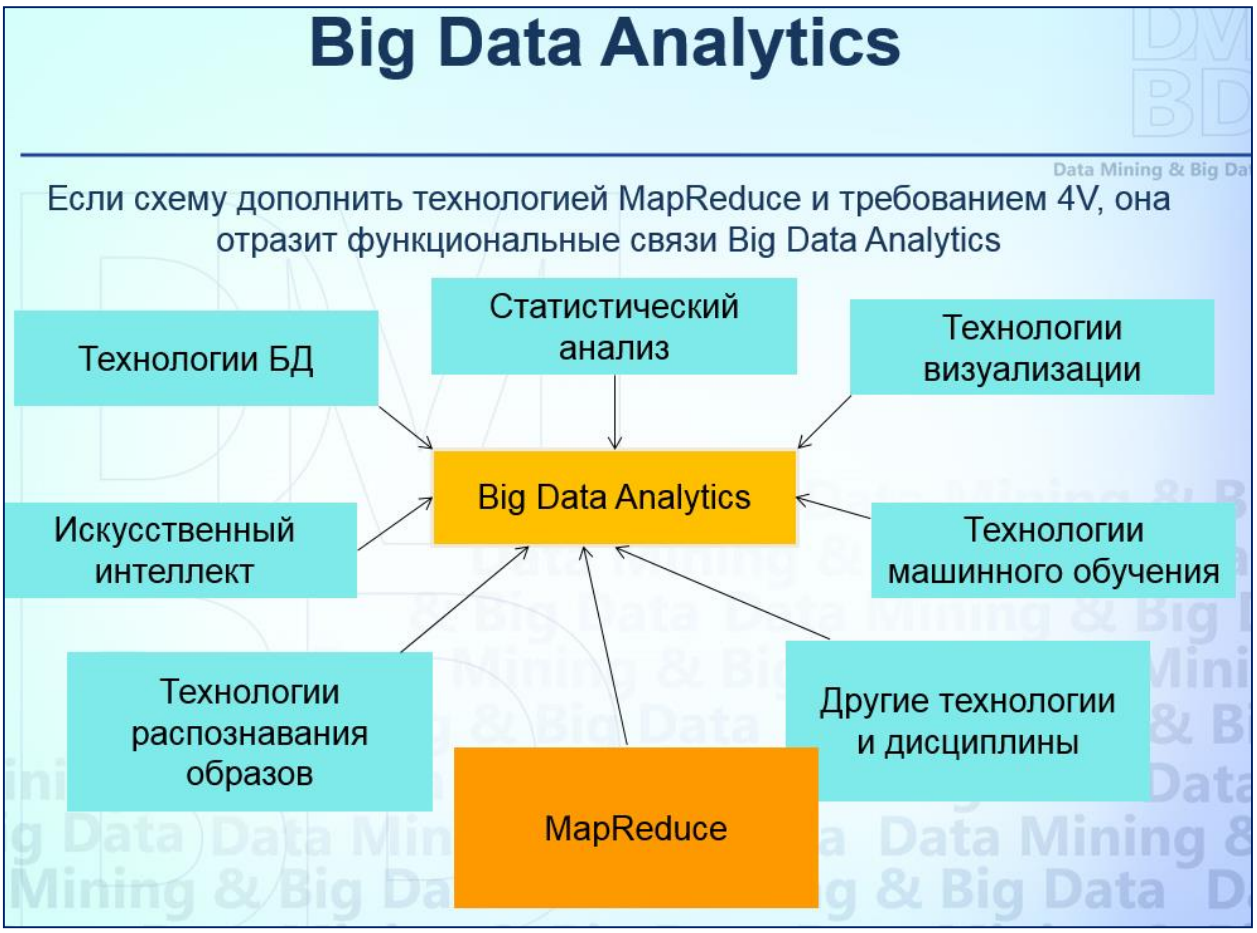
** Powered by Datameer*

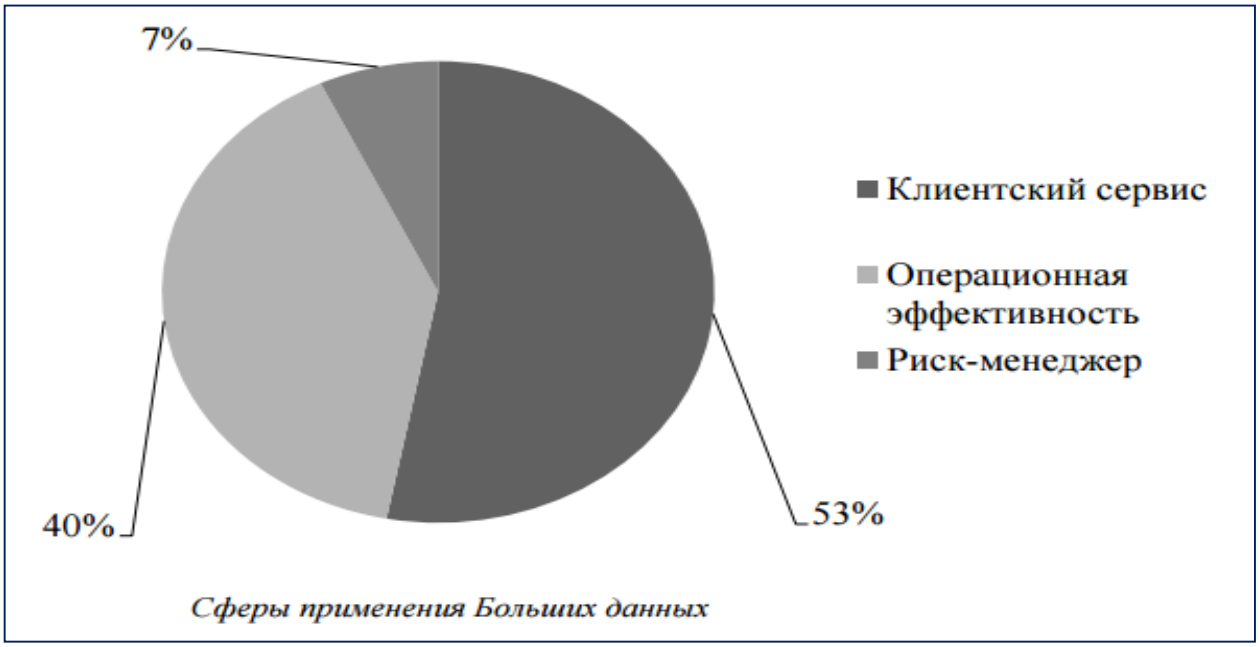
Что такое Big Data аналитика и что делает ее такой сильной?

Проблематика

До hadoop у нас было ограниченное пространство для хранения данных а также вычислительные ресурсы, что привело к длительной аналитической обработке данных(см.ниже). Во-первых, обработка проходит через длительный процесс (часто известный как ETL), чтобы получить каждый новый источник данных, готовый быть сохраненым. После того, как данные готовы, они помещаются в базу данных или хранилище данных, и в статическую модель данных. Проблема с этим подходом заключается в том, что она разрабатывает сегодня модели данных которые содержат вчерашние данные, и вы должны надеяться, что они будут актуальны для завтра.







Сферы применения Больших данных

Косвенные направления генерирования сегмента «Big Data»

Направления	Описание генерирования мультипликативных эффектов
Разработка новых видов продуктов и услуг	Это осуществляется за счет формирования более полного представления о потребностях, предпочтениях, пожеланиях покупателей, а также о них самих. В этом плане по расчетам каждое вновь созданное рабочее место в сфере «Big Data» влечет за собой возникновение 3 новых рабочих за пределами ИКТ-сектора.
Повышение эффективности хозяйственной деятельности субъектов в уже существующих отраслях	Приведем несколько примеров, например: - производитель продуктов питания Nestle за счет ИКТ-оптимизации и систематизации баз данных по своим 550 тыс. поставщикам уменьшил операционные расходы на 1 млрд. долл.; - мобильному оператору Cablecom (Швейцария) удалось снизить уровень оттока абонентов с 20 до 5%, а розничной сети Royal Shakespeare Company - увеличить число посетителей на 70% посредством развернутого анализа данных о своих клиентах.

Эффекты применения технологии «Big Data» в различных сферах

Сфера применения	Результативность применения технологии «Big Data»
Добыча полезных ископаемых	Плоскость применения помимо геологоразведки затрагивает непосредственно сам процесс добычи полезных ископаемых. Таким образом, на одном месторождении достигается сокращение операционных расходов на 10-25% и рост уровня производительности на 5%.
Государственное управление	Результируется в сокращении бюджетных расходов администрирования на 15-20%; повышение уровня собираемости налогов на 10% и увеличении эффективности государственных закупок на 30%. Также предполагается наличие действующего и масштабного по географии и спектру охвата сфер жизнедеятельности институтов электронного правительства.
Здравоохранение	Точкой опоры в данном случае выступает повседневная клиническая практика, где технологии «Big Data» могут быть использованы для существенного повышения эффективности и качества медицинского обслуживания при параллельном сокращении затрачиваемых на эти цели из государственного бюджета сумм посредством.
Наука	Пути применения их по конкретным направлениям исследований проиллюстрированы также как на примере сферы здравоохраннеия: прогностическое моделирование при разработке новых видов лекарственных средств, организация клинических испытаний на основе статистических данных, анализ тенденций заболеваемости, анализ данных клинических исследований.

Большие данные – большой успех

Согласно результатам исследования большие данные представляют значительную ценность для пользователей, реализовавших как минимум один проект. Подавляющее большинство (92%) пользователей отмечают, что они удовлетворены бизнес-результатами. Кроме того, 94% заявили, что активное использование больших данных полностью удовлетворяет их потребности. Крупные компании в большей степени ощущают на себе чрезвычайно высокую значимость больших данных для реализации их цифровой стратегии (см. рис. 1).

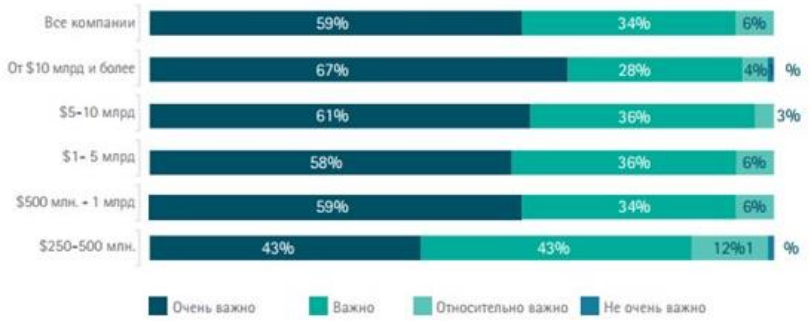
В то же время, возможно, еще многие организации пока находятся вне игры, но те, кто начал внедрять у себя большие данные, уже смогли увидеть собственными глазами практическую пользу и ощутить их ценность.

В компаниях осознают всю важность больших данных для широкого спектра стратегических корпоративных целей, начиная от поиска новых источников дохода и выхода на новые рынки и заканчивая повышением качества обслуживания клиентов и эффективности предприятия в целом.

Один из крупнейших в мире контейнерных грузоперевозчиков также входит в список крупнейших пользователей больших данных, ежегодно расходуя на хранение и обработку 16 петабайт данных, собранных со всех бизнес-подразделений, $1 миллиард.[1]

Важность больших данных

Насколько важны большие данные для вашей организации?



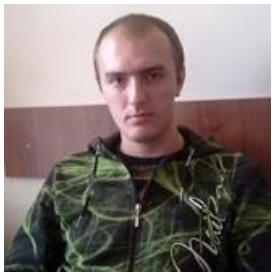
[1] Источник: http://smartdatacollective.com/bigdatastartups/201286/ why-ups-spends-over-1-billion-big-data-annually

Применение анализа больших данных

Google	**Google прекратил сообщать как много данных они хранят в 2010 (SEC filing): в то время это было 100 PBs**
YouTube	**YouTube – порядки измеряются в Exabyte** •72+ ч видео загружаются на YouTube каждую минуту •YouTube второй по использованию поисковый движок после Google •Последние данные 768+ PBs, 3-4 года назад: точно больше Exabyte сейчас
facebook	**Facebook перевалил за миллиард пользователей в августе 2012** • Население планеты стало больше 7B в прошлом году: 1/6th – в Facebook •35% мировых фотографий по оценкам в Facebook
twitter	**Twitter - около 124 млрд tweets в год, в среднем 4500 в сек**
	Обмен сообщениями в мире 193,000 смс/сек
	Звонки в США: 2.2 триллиона минут в год; 19 мин/день/человека

Заключение. Обоснованы необходимость использования и перспективность применения технологий Big Data. Приведены результаты исследований применения технологий Big Data. Исследованы современное состояние и тенденции развития технологий Big Data.

ОБРАБОТКА БОЛЬШИХ МАССИВОВ ВЫХОДНЫХ ФАЙЛОВ КОМПЬЮТЕРНОГО РЕНТГЕНОВСКОГО ТОМОГРАФА ДЛЯ РЕКОНСТРУКТИВНОЙ ЛИЦЕВОЙ ХИРУРГИИ



А.Ю. Николаев
Ассистент кафедры инженерной психологии и эргономики БГУИР, магистр технических наук, аспирант

А.Л. Раднёнок
Ассистент кафедры инженерной психологии и эргономики БГУИР, магистр технических наук, аспирант

В.С. Осипович
Доцент кафедры инженерной психологии и эргономики БГУИР, кандидат технических наук, доцент

К.Д. Яшин
Заведующий кафедой инженерной психологии и эргономики БГУИР, кандидат технических наук, доцент

Белорусский государственный университет информатики и радиоэлектроники, Республика Беларусь
E-mail: seth22@yandex.ru

Abstract. The possibility of obtaining three-dimensional computer models of the bones of the facial part of the skull and a scan of the contour of bone damage on the basis of a set of output files of an X-ray scanner.

Научный и практический интерес представляет разработка технологии для автоматизации получения трёхмерных моделей повреждений тонких, мелких костей и костей сложной формы лицевой части черепа на основе выходных файлов рентгеновского компьютерного томографа. Целью исследований явилась разработка технологии обработки больших массивов информации для реконструктивной лицевой хирургии, а также отработка технологии формирования чертежа индивидуального имплантанта костей глазницы. Все это необходимо для дальнейшего изготовления на основе этих моделей индивидуальных имплантов, замещающих повреждённые кости. Для достижения цели необходимо было решить следующие задачи: 1) автоматизировать процесс создания 3D модели повреждения; 2) автоматизировать процесс создания 3D модели импланта.

При компьютерном моделировании повреждений тонких малых костей и костей сложной формы лицевого черепа выполняют следующие операции: построение 3D модели повреждения костей; построение развёртки модели повреждения; построение рисунка будущего импланта; лазерная резка импланта; очистка и дезинфекция импланта; доводка импланта до нужной геометрии. При этом возникает следующая сложность. Тонкие кости при генерации 3D модели из выходных файлов рентгеновского компьютерного томографа пропадают или выглядят рыхлыми и дырявыми. На рисунке 1 представлен пример такой ситуации.

Такие результаты генерации 3D модели затрудняют точное определение местоположения и размеров повреждений костей. Порой совершенно не понятно, какая из костей повреждена, а какая нет.

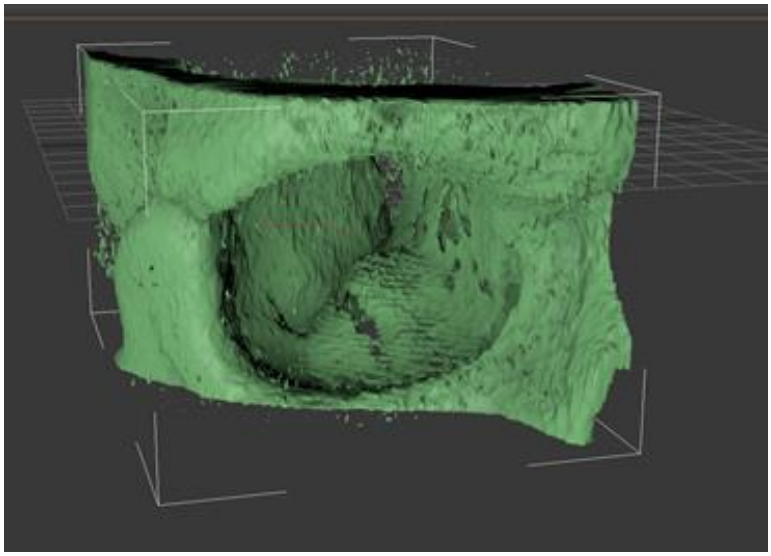


Рис. 1. Компьютерная 3D модель глазницы, построенная из выходных файлов рентгеновского томографа

Было разработано и апробировано приложение, которое обрабатывает выходные файлы рентгеновского компьютерного томографа (DICOM-файлы). При построении 3D модели костей лицевого черепа после обработки достигнуты следующие результаты.

1 При построении 3D модели лицевого черепа все здоровые тонкие и мелкие кости становятся чётко видны. Т.е. на компьютерной 3D модели будет однозначно понятно какие кости глазницы здоровы, а какие кости повреждены. На рисунке 2 представлен результат обработки выходных файлов рентгеновского компьютерного томографа с использованием разработанного компьютерного приложения.

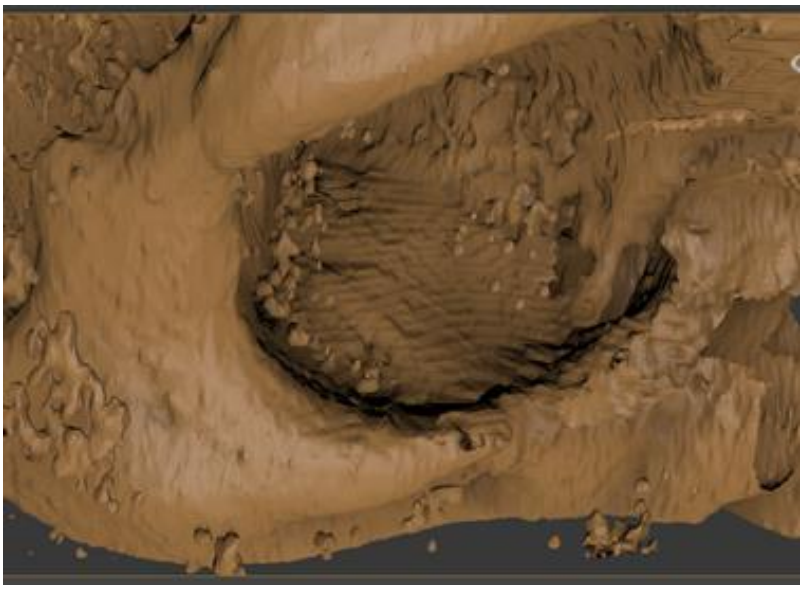


Рис. 2. 3D модель глазницы, построенная из обработанных приложением выходных файлов рентгеновского компьютерного томографа (равна модели с рисунка 1)

Можно видеть, что все мелкие отверстия в костях, которые просвечивались на исходной модели (рисунок 2), на полученной модели выглядят как сплошная кость.

12 При построении 3D модели лицевого черепа строится модель повреждения костей глазницы (геометрия поломанных костей). При обработке повреждённой глазницы приложение автоматически достраивает кость в том месте, где она должна быть и удаляет кость, которая осталась. Поэтому при построении 3D модели остаётся модель повреждения, а не здоровых костей.

Результаты построения модели повреждения говорят о том, что обработка исходных DICOM файлов в приложении позволяет получать геометрию повреждения костей в трёхмерном пространстве. Детальный анализ послойных результатов сканирования головы (DICOM файлов) показывает, что использованные алгоритмы позволяют получить более точные модели повреждений костей, в сравнении с ручной прорисовкой повреждений костей.

На рисунке 3 представлен результат совмещения модели повреждений построенных ручным способом (зелёный) и путём обработки приложением (синий). Синяя модель повреждения больше зелёной.

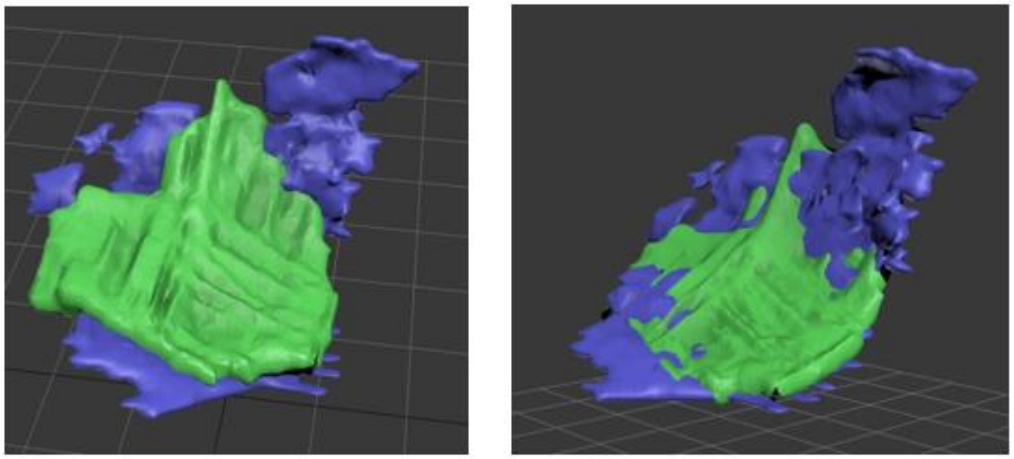


Рис. 3. 3D модели повреждений построенных разными способами

При этом алгоритм построения векторного изображения повреждения костей будет выглядеть следующим образом.1) Усиление видимости костей; 2) Создание модели повреждения; 3) Конвертация двумерной модели в трехмерную; 4) Создание развертки. На рисунке 4 представлен алгоритм построения модели импланта костей лицевого черепа.

Рис. 4. Алгоритм построения модели импланта костей лицевого черепа

Процесс получения развертки заключается в следующем. 1) Преобразование STL в обьект C#; 2) Преобразование 3D модели в 3D поверхность; 3) Построение плоской развертки 3D поверхности; 4) Формирование векторного рисунка в виде выходного файла.

Технология дает возможность создавать развертку поверхности модели повреждения кости для изготовления индивидуального импланта. Для этого необходимо провести компьютерную томографию черепа. Она определяет качество стереолитографических моделей, а они в свою очередь – соответствие индивидуальных имплантатов анатомии человека. По данным компьютерной томографии строятся трехмерные реконструкции костных структур и мягких тканей. Для создания трехмерной модели использовано свободное программное обеспечение с открытым исходным кодом, которое представляет собой гибкую, модульную платформу для анализа изображений и визуализации. На рисунке 5 представлен алгоритм построения развертки.

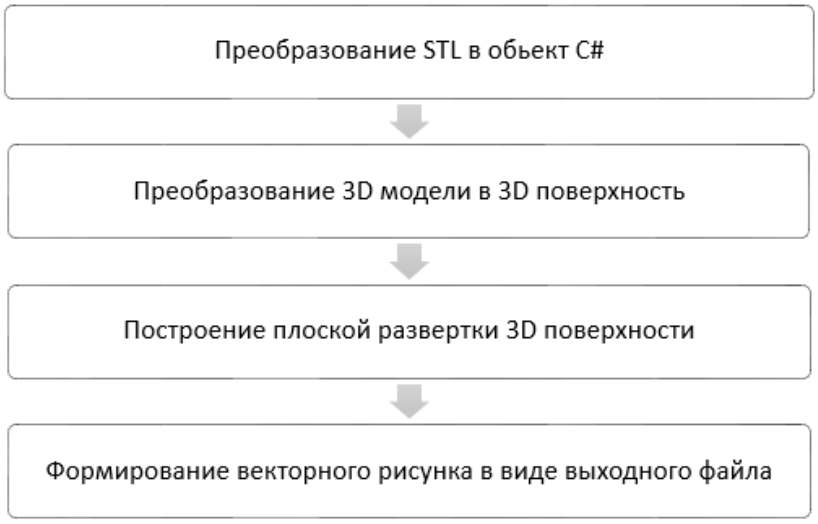


Рис. 5. Алгоритм построения развертки

Результатом технологического процесса являются файлы 3D модели костей лицевой части черепа, готовые к распечатке на 3D принтере, а также векторный файл контура повреждения костей. Изготовленная 3D модель костей лицевого черепа используется при подготовке к операции для проверки правильности изготовления импланта. Векторный файл контура повреждения костей лицевой части черепа используется для изготовления импланта. После апробации разработанной технологии распечатки модели костей лицевого черепа можно будет избежать в силу точности подготовки контура повреждения для изготовления импланта.

Литература

[1]. Анатомия головы и шеи: учебник для студ. мед. вузов / М.Р.Сапин, Д.Б.Никитюк. — М.: Издательский центр «Академия», 2010. — 336 с.

[2]. 3D Slicer [Электронный ресурс https://www.slicer.org/]

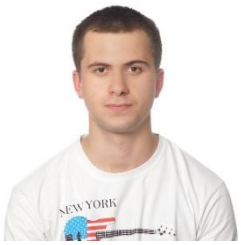
[3]. Autodesk 3D Max [Электронный ресурс http://www.autodesk.ru/]

[4]. Петцольд. Программирование для Microsoft Windows на C#. В 2-х Томах. Том 1: Пер. с англ. – Москва: Русская редакция, 2002. — 624 с.

ФУНКЦИОНАЛЬНОСТЬ СИСТЕМЫ ПОЛУЧЕНИЯ И АНАЛИЗА ТЕКСТОВЫХ ДАННЫХ

М.В. Стержанов
Ассистент кафедры информатики БГУИР

Н.Н. Шинкевич
Студентка кафедры информатики БГУИР



М.И. Селюк
Студент кафедры информатики БГУИР

Д.Н. Рожков
Студент кафедры информатики БГУИР

В.Ю. Пресняцкий
Студент кафедры информатики БГУИР

А.И. Свито
Студент кафедры информатики БГУИР

Белорусский государственный университет информатики и радиоэлектроники, Республика Беларусь
E-mail: sterjanov@bsuir.by, sn0wf1llin@gmail.com, max.selyuk@gmail.com, rdimon2912@gmail.com, presniatski@gmail.com, alexandervirk@gmail.com

Abstract. The proliferation of textual data in business is overwhelming. While the amount of textual data is increasing rapidly, businesses' ability to summarize, understand, and make sense of such data for making better business decisions remain challenging. This paper describes a system that organizes and analyzes textual data for extracting insightful customer intelligence from a large collection of documents and for using such information to improve business operations and performance.

Система представляет собой современный программный комплекс с интеллектуальной функцией анализа текстового контента, которая позволяет рассчитывать статистические характеристики текста, а также автоматически выделять ключевые слова. В разработке автоматизированной системы использованы устоявшиеся методы математического моделирования и искусственного интеллекта (в частности – аппарат искусственных нейронных сетей), математической статистики, информационных технологий и программирования.

При разработке архитектуры системы в первую очередь ставилась задача упрощения автоматического тестирования большого количества алгоритмов получения и обработки данных. Поэтому система состоит из набора отдельных слабосвязанных компонентов:

- автоматическое получение данных из внешних источников;
- подсчет статистических характеристик полученного контента;
- выделение ключевых слов.

В представленной системе используются сторонние библиотеки: NLTK WordPunktTokenizer и PunktSentenceTokenizer - для выделения предложений и слов, NLTK NaiveBayesClassifier - реализация наивного Байесовского классификатора.

Данные для обработки получаются в режиме реального времени из внешних источников, в качестве которых выступают информационно-новостные сайты.

Инструменты, позволяющие собирать данные для исследований из Веба, называются «веб-пауками» (web-spider), краулерами (web crawler) или скребками (web scraper). Поисковый робот — программный комплекс, осуществляющий навигацию по веб-ресурсам и сбор информации для базы данных приложения-агента [1]. Работу разработанного краулера можно описать следующим образом: сканирование сайта начинается с начальной страницы и затем робот использует ссылки, размещенные на ней, для перехода на другие страницы. Каждая страница сайта анализируется на наличие требуемой информации, которая копируется в соответствующее хранилище в случае обнаружения. Процесс повторяется до тех пор, пока не будет проанализировано требуемое число страниц либо пока не будет достигнута некая цель. Модуль получения данных разработан на языке программирования Ruby и состоит из трех основных частей: блок сканирования и обработки данных, блок управления краулером (команды вводятся через консоль) и база данных. Собираемая роботом информация состоит из ссылочной структуры обрабатываемого ресурса и веб-страниц. В качестве основы для базы данных была выбрана бесплатная СУБД MySQL. Для упрощения взаимодействия с БД нами используется библиотека Sequel, позволяющая представлять данные в виде объектов.

Кроме непосредственной информации, полученной с помощью краулеров, для полученного контента нами также был реализован подсчет различных статистических характеристик. Нами подсчитывается:

– общее число уникальных слов;
– общее число вопросительных предложений;
– общее число заголовков, содержащих слово Why;
– общее число заголовков, имеющих определённый контекст (проверка слова на вхождение в специальный словарь);
– общее число заголовков, содержащих числовые данные;
– наиболее часто встречающиеся слова (слово, число повторений, %);
– наиболее часто встречающиеся фразы из 2 слов (фраза, число повторений, %);
– наиболее часто встречающиеся фразы из 3 слов (фраза, число повторений, %).

В базе данных сохраняется статистика, собранная в ходе каждого запуска. В дальнейшем эта информация может быть использована как признаки (features) для алгоритмов машинного обучения.

Автоматическая классификация текста является ярким примером задач, для которых довольно сложно получить непротиворечивое, достаточно представительное обучающее множество, и в то же время, сравнительно легко собрать большой объем неразмеченных документов.

Для решения подобной задачи существует два основных подхода:

1 основанные на правилах;

2 основанные на машинном обучении - использовании статистических данных, полученных из обучающей выборки.

Подходы, основанные на правилах, в настоящее время редко используются из-за сложности, возникающей при создании правил: использование узкоспециализированных лингвистических знаний, создание ряда нетривиальных правил и невозможность обобщения результатов на другие языки.

Более перспективными являются методы машинного обучения, требующие для своей работы размеченной коллекции документов.

В данной работе в качестве коллекции документов была взята база данных статей, каждая из которых содержит название, аннотацию и содержание.

Нами было реализовано разделение коллекции документов на заданное количество тем (кластеров) с применением алгоритма латентного размещения Дирихле (Latent Dirichlet Allocation, LDA). LDA принимает набор документов, в качестве которых выступает контент

статей из коллекции документов, и выдает список тем в этих документах. Каждая тема характеризуется распределением используемых в ней ключевых слов. LDA на тренировочных данных выявляет список тем, и затем для каждой статьи можно получить вероятности отношения контента данной статьи к выявленным темам. Тема, вероятность отношения контента к которой наибольшая, выбирается в качестве темы документа. Таким образом, получается распределение статей в коллекции по темам.

Для реализации данного алгоритма использовался пакет gensim и встроенный модуль gensim.models.ldamulticore.LdaMulticore, для него был написан класс-обертка, осуществляющий подготовку текста для последующего анализа, включающий перевод текста в нижний регистр, удаление стоп-слов, пунктуации, ссылок, разбиение текста на слова, а так же задающий количество тем и слов, впоследствии используемых для описания каждой темы.

Нами реализовано выделение ключевых слов из названия, аннотации и содержания документа при помощи методов Textrank, Rake, TF-IDF для последующего сравнения и анализа.

Алгоритм Textrank основан на представлении текста в виде графа. Вершины графа – целостные части текста (отдельные слова, n-граммы, предложения). Веса дуг графа характеризуют тип связи между вершинами по выбранному принципу (например, встречаться вместе в окне размера n, т.е. на расстоянии не более n слов друг от друга). В качестве вершин графа рассматриваются отдельные слова текста; вес дуги, соединяющей две вершины-слова, показывает, сколько раз эти два слова встретились в тексте в окне n. Для оценки веса каждой вершины-слова в используется величина, основанная на модификации формулы PageRank:

$$TR(t_i) = (1-d) + d \cdot \sum_{t \in In(t_i)} \frac{w_{ji}}{\sum_{t_k \in Out(t_j)} w_{jk}} \cdot TR(t_j), \quad (1)$$

где *d* - фактор затухания, *In(t)* - множество вершин входящих в t, *Out(t)* - множество вершин исходящих из *t*, *Wij* - вес ребра *ij*.

В качестве веса ребра использовалось расстояние Левенштейна между двумя отдельными словами. В качестве результата берутся *n* слов, имеющих наибольшие значения *TR*.

Метод Rapid Automatic Keyword Extraction (RAKE) основывается на том, что ключевые слова включают в себя значимые слова, но редко включают стоп-слова, местоимения или другие слова с минимальным лексическим значением.

Извлечение ключевых слов происходит следующим образом: текст разбивается на слова по позициям стоп-слов и знаков препинания - разделителей, образуя последовательности из разделителей и собственно слов, те последовательности, которые не имеют разделителей в своем составе формируют список “кандидатов” в ключевые слова. Далее строится граф встреч данных кандидатов друг с другом в тексте документа. Вычисляется вес каждого слова как отношение

$$weight(w) = \frac{\deg(w)}{freq(w)}, \quad (2)$$

где deg(w) - word degree, freq(w) - word frequency. N слов, имеющих наибольший вес, выбираются в качестве ключевых.

Алгоритм TF-IDF[4] основан на метрике tf-idf, которая рассчитывается для каждого конкретного слова в каждом документе как произведение частоты слова в данном документе tf на инвертированную частоту документов idf, где idf определяется как

$$idf = log \frac{|N|}{df}, \quad (3)$$

где N – множество документов, df – число документов, в которых хотя бы раз встретилось слово. С помощью TF-IDF оценивается вес каждого слова в документе.

В качестве группы документов мы выбираем 300 случайных статей из коллекции.

Направлением дальнейших исследований является использование полученных статистических характеристик, выделенных ключевых слов и тем в качестве исходных данных для решения задачи предсказания различных атрибутов статей методами машинного обучения.

Литература

[1]. A.H.F. Laender, A brief survey of web data extraction tools // A.H.F. Laender et al. // ACM SIGMOD Record 31(2), pp 84-93. 2002.

[2]. Blei, D.M. Latent Dirichlet Allocation / D.M. Blei, A.Y. Ng, M.I. Jordan // Journal of Machine Learning Research. — 2003. —Vol. 3. — PP. 993 — 1022.

[3]. Agirre, Eneko and Aitor Soroa. 2009. Personalizing pagerank for word sense disambiguation. In Proc. of the 12th Conference of the European Chapter of the ACL, pages 33–41.

[4]. Aizawa, Akiko N. 2003. An information-theoretic perspective of tf-idf measures. Information Processing Management, 39(1):45–65.

СПЕКТРАЛЬНЫЙ АНАЛИЗ АСМ-ИЗОБРАЖЕНИЙ БИОЛОГИЧЕСКИХ КЛЕТОК

И.Е. Стародубцев
Аспирант БГУ

Ю.С. Харин
Директор Научно-исследовательского института прикладных проблем математики и информатики, заведующий кафедрой математического моделирования и анализа данных ФПМИ, доктор физико-математических наук, профессор, член-корреспондент НАН Беларуси

Белорусский государственный университет, Республика Беларусь
E-mail: istarodubtsev.science@gmail.com

Abstract. Spectral analysis that is a widely used method for studying big data arrays was used to analyze AFM images of biological cells such as epithelial cells line A549 (cancerous cells) and fibroblasts. Each AFM image with the resolution of 256×256 pixels was divided into 256 sections (scanning line) and for each section the periodogram was calculated. Using the obtained set of periodograms three dimensional map of spectral density estimations was formed and studied. Distinctive features of the three dimensional map of spectral density estimations for AFM images of the studied biological cells have been analyzed depending on scanning modes and temperature.

Введение. АСМ-изображение (изображение, полученное с помощью сканирующего атомно-силового микроскопа, АСМ) является массивом точек в трехмерном пространстве (x,y,z), описывающих либо карту рельефа поверхности (режим topography), либо карту локальных физико-механических свойств (режим torsion).

АСМ-изображение размером N×N пикселей можно рассматривать как совокупность из N двухмерных массивов (x,z) по N точек в каждом, расположенных на расстоянии шага сканирования вдоль оси y.

Для каждого двухмерного массива (x,z) может быть применено дискретное преобразование Фурье[1, 2]:

$$X(\omega_k) = \sum_{n=0}^{N-1}(z_n - \bar{z})e^{-j\frac{2\pi kn}{N}}, k = 1,\dots,N-1, \quad (1)$$

где $\bar{z}$ – выборочное среднее по оси z, ω_k – частота,$\omega_k = 2\pi\frac{k}{NL}$, L – максимальное значение по оси x. Затем построена периодограмма (состоятельная, но смещенная оценка спектральной плотности):

$$R(\omega_k) = |X(\omega_k)|^2, \quad (2)$$

которая позволяет построить сглаженную, асимптотически несмещенную и состоятельную оценку спектральной плотности.

Совокупность оценок спектральных плотностей N×N может быть рассмотрена как поверхность (карта), описывающая изменение спектральных характеристик АСМ-изображения вдоль оси y.

Материалы и методы. В работе был проведен анализ АСМ-изображений поверхностей биологических клеток (фибробластов и эпителиальных клеток рака лёгкого (А549) человека) в зависимости от режима сканирования и температуры сканирования образцов. Изображения размером 2.5 мкм×2.5 мкм и разрешением 256×256 пикселей, были получены в режимах сканирования topography (рельеф) и torsion (карта латеральных сил) при прямом и обратном направлениях сканирования при температурах 25°С, 50°С и 70°С для клеток А549 и 30°С и 70°С для фибробластов.

Карты спектральных плотностей были рассчитаны с помощью ПО, разработанного на языке С++ с использованием библиотеки fftw. Графики карт построены в средах Mathcad и Excel.

Результаты. На рисунке 1 видны различия карт спектральных плотностей АСМ-изображений поверхностей клеток А549 в режимах topography и torsion. «Ребра», проявляющиеся на графиках, являются, вероятно, следствием дефектов исходных АСМ-изображений (рисунок 1.1). Для АСМ-изображений поверхностей клеток А549 и фибробластов были рассчитаны средние значения оценок спектральных плотностей и построены их графики (рисунок 2). При увеличении температуры абсолютные значения оценок спектральных плотностей для клеток А549 и фибробластов в режиме torsion увеличиваются (рисунки 1.2, 1.4, 2). На исходных АСМ-изображениях клеточной поверхности в режиме topography присутствует более гладкая локальная область, и она отчетливо проявляется на карте спектральных плотностей в виде области более низких значений (рисунки 1.1, 1.3, 2).

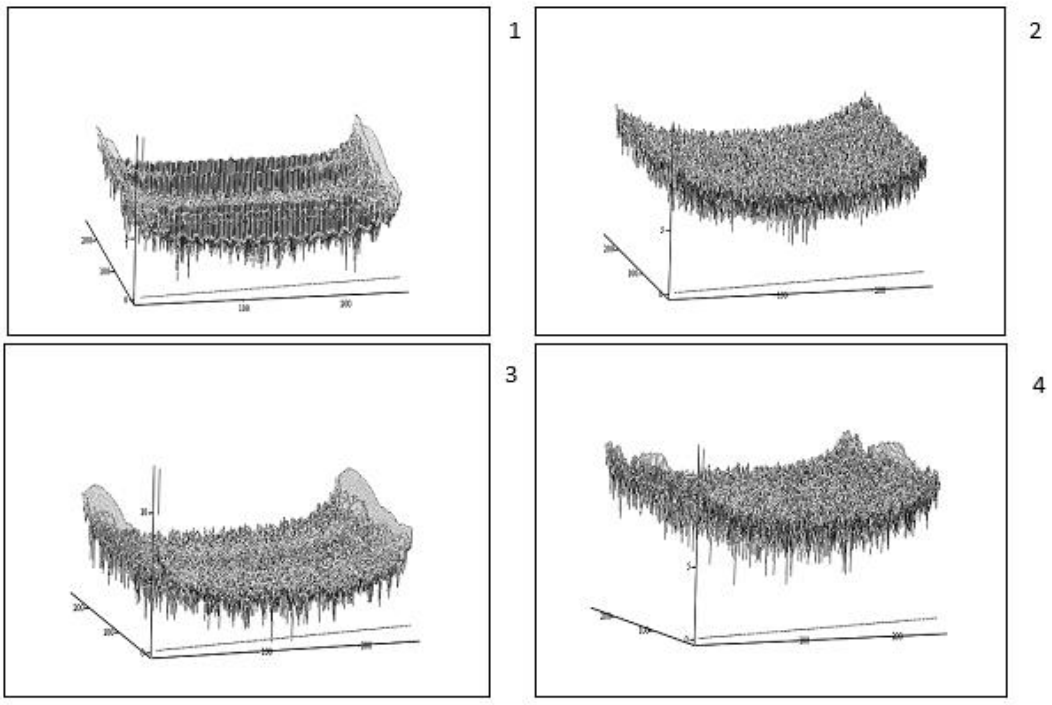
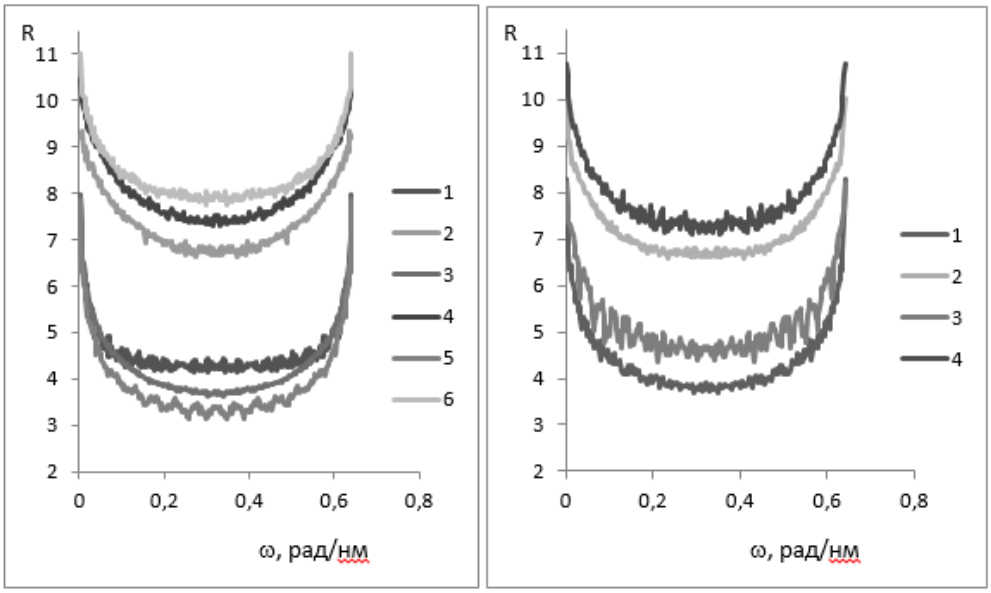


Рис. 1. Карты спектральных плотностей для АСМ-изображений клеток А549, полученных в режимах сканирования topography (1,3) и torsion (2,4) и температурах 25°С (1,2) и 70°С (3,4).

Для численной оценки карты спектральных плотностей были разбиты по оси x (частота, ω) на 7 равных фрагментов и для каждого фрагмента была рассчитана фрактальная размерность (D_F) методом подсчета кубов (box counting)[3] (рисунки 3, 4).

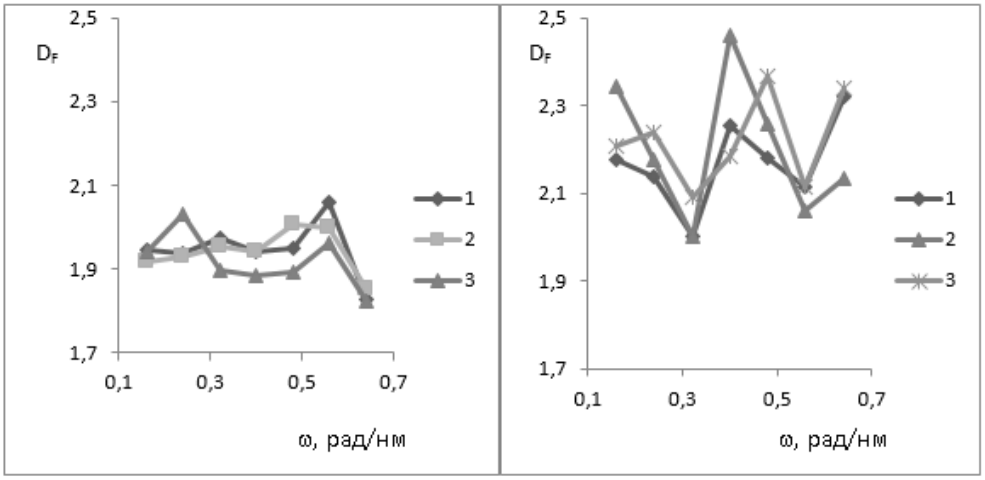
Анализ частотной зависимости фрактальной размерности для клеток А549 выявил, что

наиболее сложные структуры характерны для области частот 0.5-0.6 рад/нм в режиме topography и области частот 0.4-0.5 рад/нм в режиме torsion (рисунок 3). Для фибробластов частотные зависимости фрактальной размерности в обоих режимах также различается (рисунок 4).



А Б

Рис. 2. А: средние значения спектральных плотностей для АСМ-изображений клеток А549, полученных в режимах сканирования topography (1,3,5) и torsion (2,4,6) и температурах 25°С (1,2), 50°С (3,4) и 70°С (5,6). Б: средние значения спектральных плотностей для АСМ-изображений фибробластов, полученных в режимах сканирования topography (1,3) и torsion (2,4) и температурах 30°С (1,2) и 70°С (3,4)



А Б

Рис. 3. Зависимость фрактальной размерности от частотного диапазона для: АСМ-изображений поверхности раковой клетки А549, полученных в режимах сканирования topography (А) и torsion (Б) при температурах 25°С (1), 50°С (2) и 70°С (3).

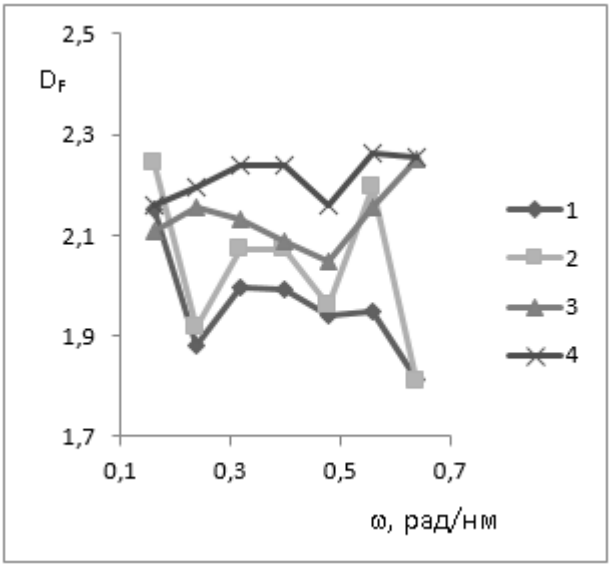


Рис. 4. Зависимость фрактальной размерности от частотного диапазона для: АСМ-изображений поверхности фибробласта, полученных в режимах сканирования topography (1, 2) и torsion (3, 4) при температурах 30°С (1, 3), и 70°С (2, 4).

Для обоих типов клеток значения фрактальной размерности карт спектральной плотности для режима torsion (2.00-2.46 рад/нм) больше, чем наблюдаемые для режима topography (1.79-2.09 рад/нм). Это свидетельствует о том, что карты оценок спектральной плотности для режима torsion имеют в целом более сложную структуру в сравнении с со структурой карт для режима topography.

Заключение. Предложная методика, включающая построение карт оценок спектральных плотностей с последующим расчетом фрактальной размерности в различных частотных диапазонах, применима для анализа изменений структуры и свойств поверхностей биологических клеток по АСМ-изображениям.

Литература

[1]. Харин Ю. С. Теория вероятностей, математическая и прикладная статистика: учебник / Ю. С. Харин, Н. М. Зуев, Е. Е. Жук. - Минск: БГУ, 2011. - 463 с.

[2]. Julius O. Smith III. Mathematics of Discrete Fourier Transformation (DFT) with audio applications.- W3K Publishing, 2007. - 322 p.

[3]. Novel fractal characteristic of atomic force microscopy images / M.N. Starodubtseva, I.E. Starodubtsev, E.G. Starodubtsev. // Micron. -2017. –Vol. 96. –P. 96-102

ИНФОРМАЦИОННЫЕ МОДЕЛИ ПСИХОЛОГИЧЕСКОГО ВЛИЯНИЯ РЕКЛАМЫ НА ПОТРЕБИТЕЛЯ

Л.А. Вайнштейн
Профессор кафедры инженерной психологии и эргономики БГУИР, кандидат психологических наук, доцент

Белорусский государственный университет инфоматики и радиоэлектроники, Республика Беларусь

Abstract. The analysis of various information models used in advertising. The role of various psychological mechanisms and mental processes advertising effect on the user. Set out a comparative analysis of the effectiveness of different infor-mation

В настоящее время реклама является фактором, широко используемым для влияния на поведение потребителей. Одним из наиболее используемых видов рекламы является представление людям различной визуальной информации по разнообразной продукции. Причем, для этого все шире используется интернет реклама.

Исторический анализ позволяет выделить две традиции или две теоретические тенденции в развитии психологии рекламы. Первая возникла на основе многочисленных экспериментальных психологических работ в области психических процессов в рекламе, вторая – на основе развития идей маркетинга. Первая традиция наибольшего расцвета достигла в первой половине XX в. в Германии (ее условно можно назвать немецкой) и рассматривала в качестве основного – фактор социального воздействия с учетом закономерностей переработки информации. Вторая – возникла в США во второй половине XX в. (ее условно можно назвать американской) и рассматривала потребности людей и процесс их «опредмечивания» рекламой.

Однако, несмотря на имеющиеся традиции и широкое применение различной рекламы, она часто далека от необходимого информационного воздействия на человека из-за игнорирования и неучета психологических факторов воздействия визуальной информации. В результате у людей не всегда формируется правильные установки действия, что снижает эффективность применения рекламы. Человек «вроде видит», но не воспринимает представленную рекламную информацию, особенно, отображенную на сайте компьютера или планшета.

Причин здесь может быть несколько. Во-первых, неудачное решение дизайна сайта без учета требований эргономики, во-вторых, неудачное юзабилити сайта из-за расположение информационных элементов и алгоритма пользования.

Много лет назад в психологии возникло направление, целью которого является изучение структуры психологического воздействия на человека визуальной информации. Это направление получило большое распространение, в частности, в психологии рекламы и широко используется в западной и российской практике. Представление визуальной информации человеку основано на базовых психологических законах информационной деятельности человека (рис.1), поэтому разработанные сейчас психологические модели могут быть применены в рекламе для психологического визуального воздействия, представляемой информации [1].

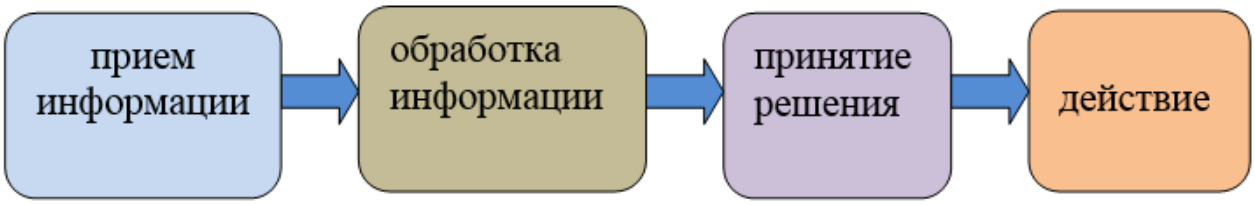


Рис.1. Структура информационной деятельности человека

Рассмотрим основные модели процессов визуального воздействия информации на человека [2,3]. Следует указать, что еще в конце XIX в. предпринимались попытки разработать некую обобщенную теоретическую модель, описывающую наиболее эффективную структуру информационного воздействия.

Одной из первых появилась теоретическая модель, основанная на формуле AIDA, которая была предложена Элмером Левисом в 1896 г. Автор считал, что воздействие информации всегда начинается с *привлечения внимания (attention),* затем она должна вызвать *интерес (interest),* потом *желание (desire)* и после этого, как правило, *возникает активизация деятельности (activity).*

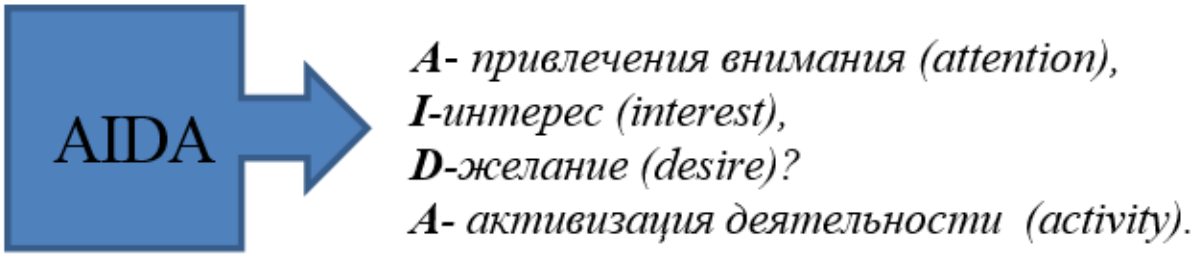


Рис.2.Воздействие информации на человека по формуле AIDA (англ.)

Внимание — избирательная направленность восприятия на тот или иной объект. Изменение внимания выражается в изменении переживания степени ясности и отчётливости предмета деятельности человека. Внимание находит себе выражение в отношении человека к объекту. За ним стоят интересы и потребности, установки и направленность человека. Это, прежде всего, вызывает изменение отношения к объекту, выражаемое вниманием — его сознаваемостью. Внимание обуславливает успешную ориентировку субъекта в окружающем мире и обеспечивает более полное и отчётливое отражение его в психике. Объект внимания оказывается в центре нашего сознания, все остальное воспринимается слабо, неотчётливо, однако направленность нашего внимания может меняться.

Интерес – положительно окрашенный эмоциональный процесс, связанный с потребностью человека узнать что-то новое об объекте интереса, повышенным вниманием к нему. Мотив – динамический процесс физиологического и психологического плана, управляющий поведением человека, определяющий его направленность, организованность, активность и устойчивость. Часто определяется как «опредмеченная потребность». Желание – потребность, принявшая конкретную форму в соответствии: с культурным и профессиональным уровнем и личностью человека. Деятельность — целеустремленная активность, реализующая потребности человека.

Позже в формулу был внесен еще один элемент — мотив (motive). Формула приобрела окончательный вид — AIMDA, где A- внимание, I - интерес,

M- мотив, D - желание, A - активизация деятельности.

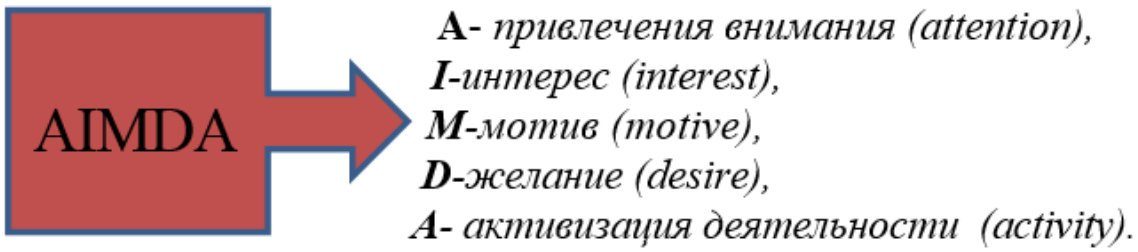


Рис.3. Воздействие информации на человека по формуле AIMDA(англ.)

Однако в этой теоретической модели не до конца было установлено, как взаимодействуют между собой элементы данной схемы. Например, каким образом, начинаясь от привлечения внимания, информационное воздействие на человека заканчивается его исполнительской деятельностью. То есть формула не до конца раскрывала связи между информационным воздействием и заранее предполагаемым результатом деятельности. Очевидно, что возможность приобретения человеком какого-либо товара не является прямым следствием привлечения внимания, как следует из данной формулы и как иногда ошибочно полагают специалисты. Между привлечением внимания и поступком существует сложная цепь причинно-следственных связей, определяемых достаточным количеством психологических и других факторов. Кроме того, как показала практика, в данную формулу не попали такие важные переменные, как потребности и мотивы человека, память, эмоции, ассоциативное мышление, социально-психологические установки и др., которые играют очень важную роль в процессе принятия решения под воздействием воспринимаемой информации.

В настоящее время накоплен большой опыт применения данной теоретической модели с использованием формулы AIMDA, хотя большое число специалистов сегодня ее критикуют.

Несколько позже была предложена формула ACCA, которая характеризуется тем, что сводит эффект информационного воздействия к определению аудитории, прошедшей через один из четырех этапов потребительского поведения — внимание *(attention),* восприятие аргументов *(comprehension),* убеждение *(conviction)* и действие *(action).*

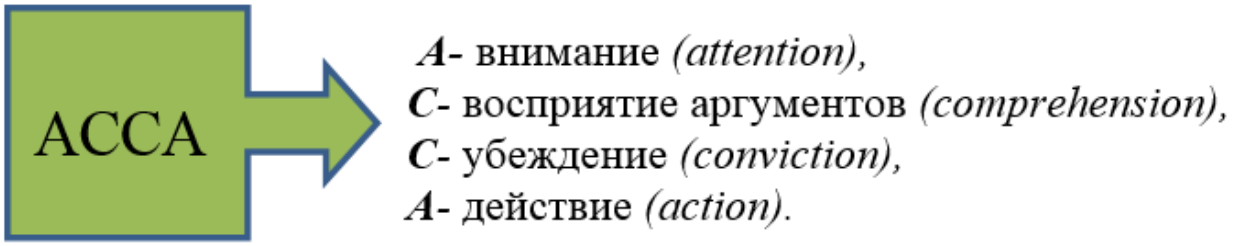


Рис.4. Воздействие информации на человека по формуле ACCA (англ.)

В качестве одного из основных элементов психологического воздействия данная модель включает убеждение. Известно, что убеждение является одним из наиболее действенных механизмов воздействия на стратегию на мышление человека. В то же время особенностью данной модели является недооценка роли потребностей человека в структуре информационного визуального воздействия.

Ведь человека нельзя убедить или заставить захотеть, делать что- либо, в чем у него изначально нет объективной нужды и потребности в этом. Очевидное достоинство формулы — внимательное отношение к процессу мышления. При разработке информационных средств на основе этой формулы делается акцент на мышление работника, на его осознанное поведение в условиях производственных рисков.

На основе дальнейших исследований (Г. Гольдман) была предложена формула DIBABA. Она основана на аббревиатуре немецких названий этапов процесса принятия решения человеком: 1) определение потребностей и желаний потенциальных потребителей; 2) отождествление потребительских нужд с предложением рекламы по охране труда; 3) «подталкивание» потребителя к необходимым выводам о действиях, которые ассоциируются с его потребностями; 4) учет предполагаемой реакции потребителя; 5) вызов у потребителя желания действовать; 6) создание благоприятной для действия обстановки.

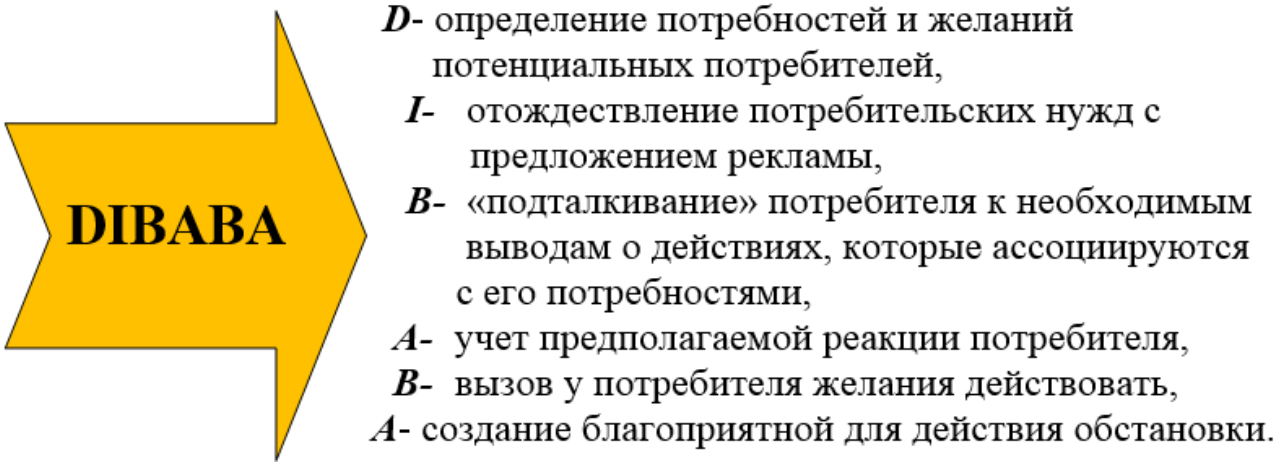


Рис.5. Воздействие информации на человека по формуле DIBABA (нем.)

Преимуществом данной модели являются: ориентация на потребности в информации человека, понимание роли процесса принятия решений, сравнения, осознанного выбора в той или иной ситуации, использование законов мышления, введение в модель механизма принятия решения и «обратной связи», понимание роли эмоций и позитивного отношения человека к рекламе.

Новый этап в разработке психологической структуры информационного воздействия в виде краткой формулы был продолжен американским рекламистом Расселлом Колли, который предложил модель DAGMAR. Несмотря на то, что она предложена автором применительно к рекламе продукции с целью повышения продаж, данная формула, как показала практика, может применяться и для других случаев визуального информационного воздействия на человека.

Формула образуется из начальных букв английской фразы: *Defining advertising goals — measuring advertising results* (определение рекламных целей — измерение результатов рекламы). Согласно этой модели, акт покупки проходит четыре фазы: 1) узнавание марки товара; 2) ассимиляция — осведомление адресата о качестве товара; 3) убеждение — психологическое предрасположение к покупке; 4) действие — совершение покупки адресатом рекламы.

Эффект рекламы определяется приростом числа покупателей на каждой из указанных фаз. Отличие модели DAGMAR от подходов, ориентированных на действия, заключается в исходной посылке: совершение действия определяется всеми основными элементами комплекса маркетинга.

В последнее время в качестве концепции психологического воздействия информации на человека используется модель социально-психологической установки, или аттитюда (*attitude).* Данная модель предполагает, что в процессе информационного воздействия у субъекта возникает готовность к действию, которая имеет сложную многокомпонентную структуру. Выделяют следующие компоненты: *познавательный* (когнитивный), *эмоциональный* (аффективный), *поведенческий* (конативный) (рис.6).

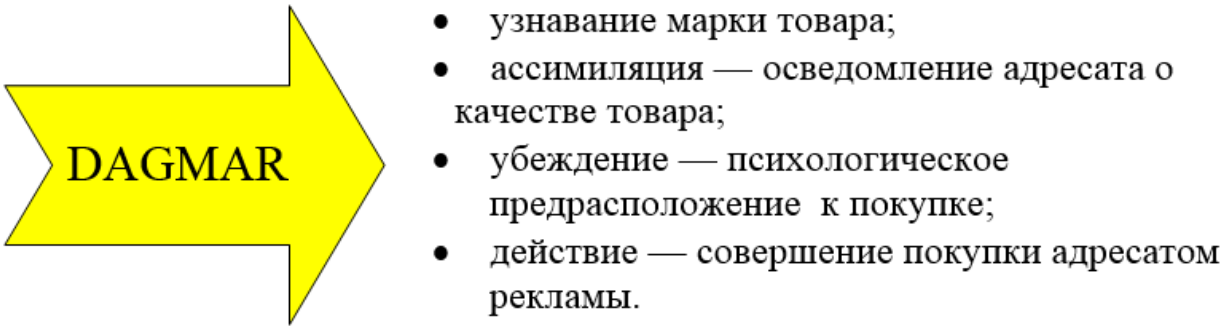


Рис.6. Воздействие информации на человека по формуле DAGMAR (англ.) Defining advertising goals — measuring advertising results (определение рекламных целей — измерение результатов рекламы).

При информационном воздействии у субъекта возникает готовность к действию, имеющего сложную многокомпонентную структуру.

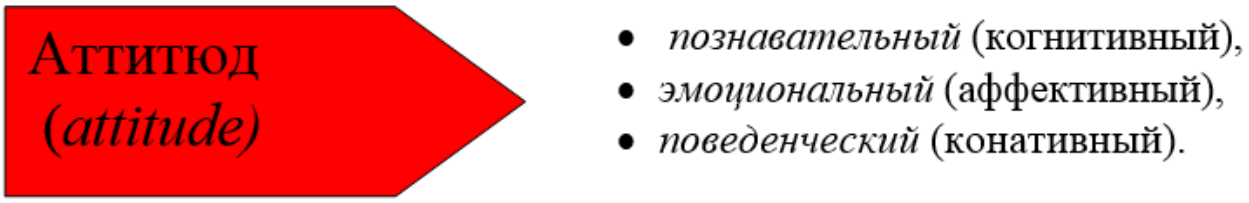


Рис.7. Воздействие информации на человека по модели аттитюда (attitude)

Предполагается, что социально-психологическая установка оказывается эффективной для поведения, если между ее отдельными компонентами нет существенных противоречий. Преобладание одного компонента над другими приводит к ослаблению воздействия получаемой установки, к снижению степени ее влияния на поведение человека. Достоинством данного подхода можно считать попытку, когда используя систему психологических понятий, можно подвергнуть научному анализу максимальное количество участвующих в нем психических процессов. Так, познавательный компонент предполагает анализ следующих процессов переработки информации:

–восприятия, внимания, памяти, принятия решений, прогнозирования, планирования, мышления;

–эмоциональный компонент — анализ эмоциональных состояний, отношений и др.;

–конативный — анализ поступков, неосознаваемых факторов, влияющих на эти поступки и т. д.

Однако в направлении, развиваемом только на основе теории аттитюда, не учитывается определяющая роль объективных потребностей человека как основного фактора поведения.

В последнее время в психологии личности в качестве объяснительного принципа стали чаще использовать модели, в основу которых положены ситуативные факторы поведения потребителей. Одним из ярких примеров является теория «базиса отсчета» Музафера и Кэролин Шерифов. По нашему мнению данная концепция хорошо учитывает поведение потребителей в зависимости от возникающих ситуаций.

Традиционные психологические теории, исследующие мотивацию и поведение людей, основываются на абстрактных понятиях, например, говорят, что у человека есть некоторая потребность, и он действует в соответствии с ней. Специфика данной теории состоит в том,

что она рассматривает поведение человека в максимально конкретных условиях в данный момент времени. Такое поведение возникает из «психологического настроя», который может быть «схематизирован». Иначе говоря, представляет собой «приказ» индивиду на обработку конкретного комплекса раздражителей, определяющих поведение человека. Правда в этом случае со стороны бывает иногда сложно предсказать или объяснить причины конкретного поступка человека. Например, почему один человек в процессе покупки руководствуется преимущественно ценой, а другой качеством изделия. Последнее особенно проявляется в приобретении технически сложных изделий долговременного пользования.

Объяснить это можно следующим. Факторы, определяющие психологический настрой в любой момент времени, могут быть различными и делятся на «внешние» и «внутренние». К «внешним» относятся – люди, погодные условия, технологии и др., а к «внутренним» — что происходит в этот момент времени внутри человека. Это могут быть потребности, мотивы, воспоминания, отношение к чему-либо, состояние здоровья и т. д.

Каждый человек, в соответствии с теорией «базиса отсчета», сознательно или бессознательно постоянно выбирает некоторые из этих внутренних и внешних факторов и при этом игнорирует другие. Способность к отсеиванию тех или иных факторов время от времени может меняться. С течением времени селективность человека начинает превращаться в определенную схему. Он приобретает склонность отдавать предпочтение одним вещам в противовес другим. Возникают так называемые «якоря» (как аналог психологической установки). Это происходит потому, что человеку от природы дано схематизировать опыт, что, по-видимому, объясняется его желанием избежать информационной перегрузки [3].

Однако чаще всего стимулов так много и они так разнообразны, что схематизации, стабилизации может не наступить. Так, часто наблюдается, что человеку сложно прийти к какому-то однозначному решению, т. е. сделать выбор. Ряд людей часто думают, что сделать выбор должны не они, а продавцы, продающие товар. При этом, они забывают, что в качестве пострадавших оказываются в первую очередь они сами, когда приобретают не тот товар. Здесь информация, предлагая четкую рекомендацию, «наводит порядок» в нестабильном поле значимых и часто противоречивых факторов. Именно этим и удается воздействовать на поведение человека, нередко вынуждая его принимать то решение, которое выгодно продавцу.

Однако реакция человека зависит в основном от того, насколько схематизация, заданная разработчиком рекламы, соответствует представлениям человека как покупателя. В случаях, когда внешний раздражитель структурно оформлен достаточно четко, влияние внутренних факторов (например, способность видеть то, что хочется увидеть) ослабляется.

Теоретические разработки в области психологии информационного воздействия в настоящее время позволяют определить формальные границы, в пределах которых эта деятельность оказывается эффективной. Главное, они помогают ответить на вопрос, что происходит с человеком, когда под воздействием информации он принимает решение о конкретном действии в той или иной ситуации.

Как показывает практика, реально, при принятии решений человеком обнаруживается огромное количество факторов, которые оставляют за ним право действовать осознанно и не превращают в пассивное существо, полностью зависимое от внешних, часто неопределенных условий. Знания различных моделей сложных психологических процессов позволяет объяснять поведение потребителя, разворачивающиеся в конкретных условиях. Поэтому сопоставление понятий и категорий, установление функциональных и предметных связей между ними сегодня является актуальной задачей психологии рекламного воздействия.

В противном случае возникают проблемы взаимопонимания между разработчиками рекламы и потребителями, появляются ситуации, препятствующие достижению механизмов информационного воздействия. В результате затраченные финансовые средства на разработку визуальной информации используются формально и неэффективно, принося вместо прибыли - убытки.

Таким образом, знание психологических механизмов психологического воздействия визуальной информации позволяет научно обоснованно использовать различные подходы в разработке различной рекламы, что позволяет повысить её эффективность.

Литература

[1]. Вайнштейн Л.А. Эргономика: учебное пособие /Л.А. Вайнштейн. – Минск: ГИУСТ БГУ, 2010.

[2]. Вайнштейн Л.А. Экономическая психология. – Минск: БГУ, Электронная библиотека БГУ, 2011.

[3]. Вайнштейн Л.А. Психология восприятия. /Л.А. Вайнштейн. – Минск: Тессей, 2007.

ПРИМЕНЕНИЕ КОМПЛЕКСНОГО ПОДХОДА К ВЕБ-АНАЛИТИКЕ

И.Ф. Киринович
Доцент кафедры инженерной психологии и эргономики БГУИР, кандидат физико-математических наук, доцент

А. А. Белов
Аспирант кафедры инженерной психологии и эргономики БГУИР

Белорусский государственный университет инфоматики и радиоэлектроники, Республика Беларусь
E-mail: artsem.bialou@gmail.com, Kirinovich@bsuir.by

Abstract. A comprehensive approach to web analytics is described, which can be presented in different data sections and must be dynamically connected and configurable. Includes ways to create an analytics profile through a code snippet, as well as using the Tag Manager.

В данной работе описан комплексный подход к веб-аналитике, которая может быть представлена в различных разрезах данных и должна быть динамически подключаемой и конфигурируемой.

В общем случае веб-аналитика позволяет оценить эффективность веб-ресурса и улучшить его работу, в том числе увеличить количество посетителей и повысить уровень продаж [1]. Таким образом, веб-анализ превращается в набор способов и инструментов, которые помогают выявить проблемы, критически подойти к работе сайта и оценить его функциональность. Однако аналитика не является исключительно теоретическим методом, это также комплекс мер, которые направлены на улучшение ресурсов. Процесс аналитического исследования не должен ограничиваться каким-либо этапом разработки системы. Это продолжительный во времени процесс выявления и определения показателей работы системы.

На смену устаревшим технологиям приходят более эффективные решения, в том числе модули веб-аналитики. Поэтому возникает необходимость динамического подключения новых модулей либо отключения существующих (из-за неэффективности, устаревания, стоимости поддержки, скорости работы).

Модуль аналитики (некоторая библиотека) выполняется на странице сайта: регистрирует события, отслеживает пользовательские действия и т.д. Так как для отображения данных используется HTML, то библиотеки разрабатываются на языке JavaScript и исполняются интерпретатором браузера. Например, на сайте Google Analytics [2] после регистрации и создания профиля аналитики для конкретного ресурса генерируется участок кода, который должен быть вставлен на страницу для отправки статистических данных, как отображено на рисунке 1:

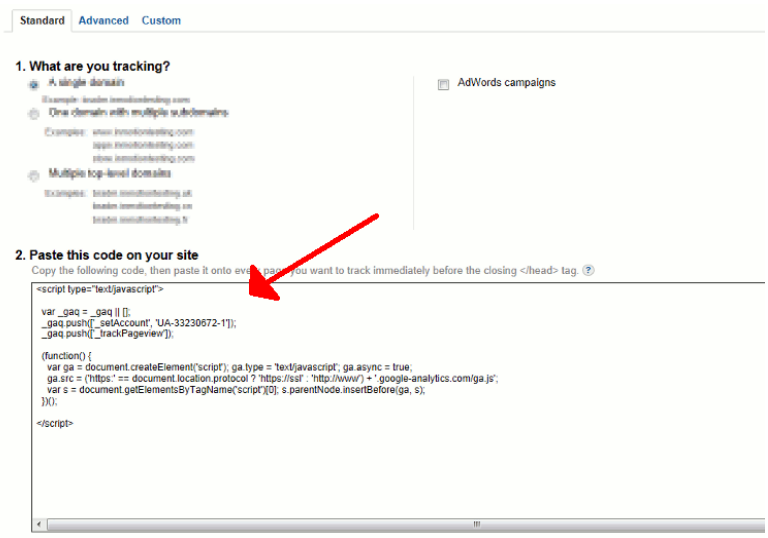


Рис. 1. Подключение модуля аналитики к сайту

Таким образом, изменение конфигурации аналитики на сайте требует замены фрагментов кода, отвечающих за отправку статистических данных. Данными фрагментами кода удобно управлять внешне, как предлагает практика внедрения зависимостей [7]. Существует готовое решение от Google (диспетчер тегов), позволяющее решить данную задачу в режиме онлайн без перезагрузки системы, а также без изменения кода её компонентов.

Диспетчер тегов Google – это система управления тегами, позволяющая быстро обновлять теги и фрагменты кода на сайте или в приложении, добавлять и изменять теги AdWords, GoogleAnalytics, FirebaseAnalytics, Floodlight, а также сторонние и пользовательские теги без внесения изменений в код сайта. Благодаря этому происходит значительная экономия времени, сокращается количество ошибок и отпадает необходимость обращаться за помощью к разработчику [3]. На рисунке 2 представлен список наиболее распространённых тегов, поддерживаемых тег-менеджером Google.

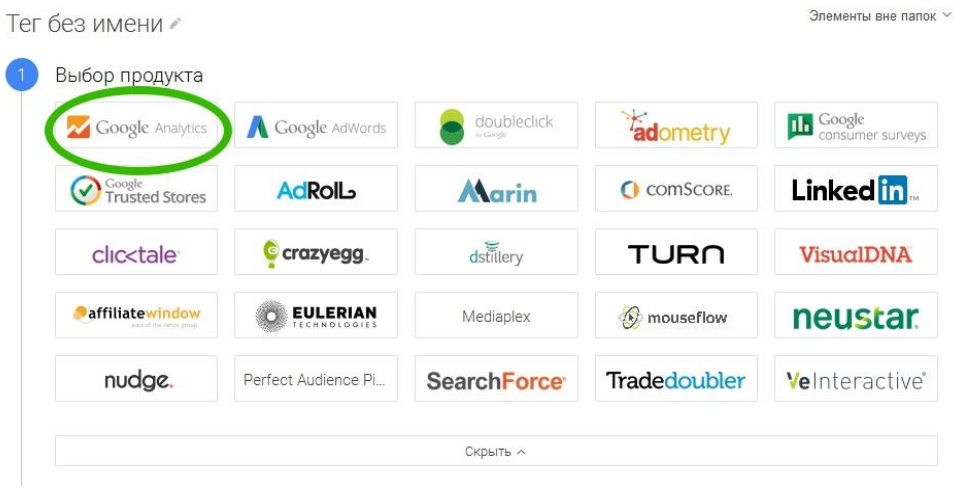


Рис. 2. Поддерживаемые теги диспетчером тегов Google

Тег представляет собой фрагмент кода, который собирает и отправляет данные с вебсайтов и приложений третьей стороне. Теги можно вставлять в исходный код сайта, а также мобильного приложения вручную либо с помощью специального инструмента. Благодаря Диспетчеру тегов, достаточно указать в интерфейсе теги, которые необходимо использовать,

а также время их активизации.

В диспетчере тегов используется тег-контейнер, который необходимо разместить на всех страницах сайта. Контейнер заменяет все теги, вручную добавленные в код сайта или приложения. Разместив тег-контейнер, можно добавлять и обновлять теги, а также управлять их работой непосредственно в интерфейсе Диспетчера тегов.

С помощью Диспетчера тегов можно управлять тегами одного или нескольких веб-сайтов и мобильных приложений. Для аккаунта Google можно создать несколько аккаунтов Диспетчера тегов, однако, в большинстве случаев достаточно одного.

При добавлении слоя веб-аналитики на веб-ресурс, придерживаются определённой последовательности действий. После изучения структуры веб-ресурса, необходимо выбрать данные, которые коллекционируются, сконфигурировать сущности, используемые для отправки статистических данных, настроить диспетчер тегов. Последовательность действий может быть такая, как показано на рисунке 3.

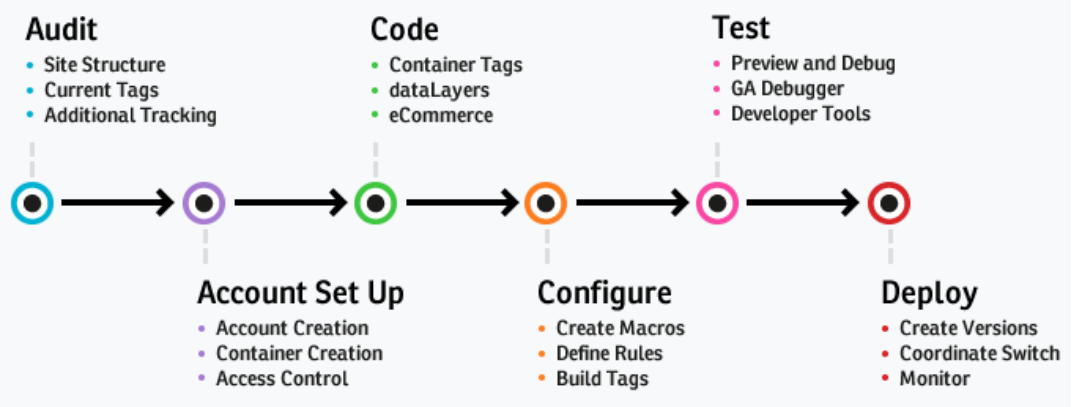


Рис. 3. Последовательность действий по добавлению аналитики в веб-ресурс [7]

Диспетчер тегов Google приведён как пример, теоретически это может быть другой диспетчер тегов. Что касается использования данных, которые коллекционируются, то их использование также варьируется. Они могут использоваться для построения графиков, таблиц, тепловых карт активности и т.д..

Графики и таблицы удобно создавать с помощью Google Analytics либо другого продукта. В рассмотрение берут не только набор функциональности, предоставляемый сервисом, а также коммрческую выгоду от использования того или иного продукта.

Тепловые карты активности можно строить с помощью сервиса Clicktale. Теги Clicktale могут быть добавлены на веб-ресурс без диспетчера тегов, однако при использовании вышеописанного подхода можно получить дополнительные преимущетсва, не в ущерб функциональности сервиса. На основе собранных данных Clicktale сервис позволяет выявлять наиболее востребованные функции сайта и отображать их на тепловой карте активности пользователя [5]. Командой разработчиков навигация к таким элементам упрощается, либо компоненты перемещаются выше на странице (во избежание скроллинга). Таким образом, повышается удобство использования ресурса. На рисунке 4 представлен пример тепловой карты страницы сайта:

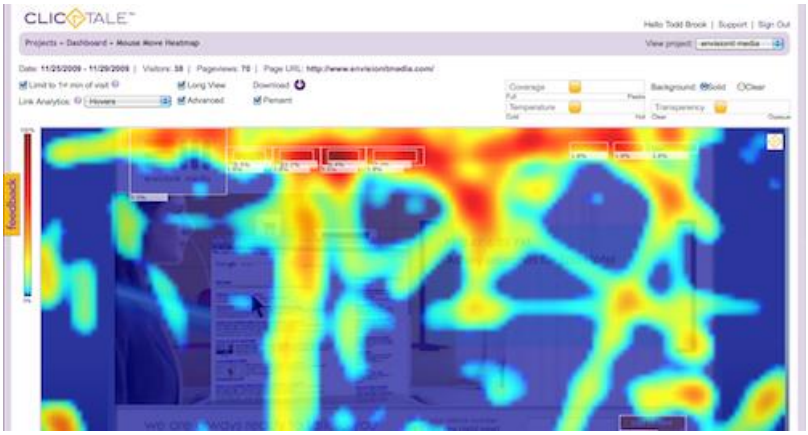


Рис. 4. Тепловая карта активности пользователя на странице сайта

Диспетчер тегов удобно применять и для A/B-тестирования. A/B-тестирование – метод исследования, суть которого заключается в том, что контрольная группа элементов сравнивается с набором тестовых групп, в которых один или несколько показателей были изменены, для того, чтобы выяснить, какие из изменений улучшают целевой показатель [4]. Разновидностью A/B-тестирования является многовариантное тестирование. В этом случае тестируются не два целостных варианта, а сразу несколько элементов продукта или составных частей исследуемого объекта в различных сочетаниях, при которых каждый тестируемый элемент может быть двух видов (A или B). Для каждого варианта можно добавить параметр, указывающий источник данных. Затем на формах аналитике по данному параметру можно идентифицировать различные варианты и построить сравнительную диаграмму в разрезе необходимых статистических данных.

Литература

[1]. Веб-аналитика [Электронный ресурс]. – Электронные данные. – Режим доступа https://ru.wikipedia.org/wiki/Веб-аналитика.

[2]. Консоль управления Google Analytics [Электронный ресурс]. – Электронные данные. – Режим доступа https://analytics.google.com/analyti cs/web/#report/defaultid/a64380901w100310836p104194387/.

[3]. Диспетчер тегов Google [Электронный ресурс]. – Электронные данные. – Режим доступа https://support.google.com/tagmanager/answer/ 6102821?hl=ru.

[4]. A/B-тестирование [Электронный ресурс]. – Электронные данные. – Режим доступа https://ru.wikipedia.org/wiki/A/B-тестирование

[5]. Click Tale official [Электронный ресурс]. – Электронные данные. –Режим доступа http://www.inspectlet.com/?gclid=Cj0KEQjw9vi-BRCx1_GZgN7 N4voBEiQAaACKVgxs1xArmRRe14-XN6HO-zy8ryF-cHdlDIzxGonw3b4aAp9j8P8HAQ.

[6]. Dependency injection [Электронный ресурс]. – Электронные данные. – Режим доступа https://en.wikipedia.org/wiki/Dependency_injection.

[7]. Tag Manager 360 [Электронный ресурс]. – Электронные данные. – Режим доступа http://www.periscopix.co.uk/analytics-360-suite/tag-manager-360.

МОДЕЛИРОВАНИЕ РИСКА БАНКРОТСТВА ПРЕДПРИЯТИЙ РЕАЛЬНОГО СЕКТОРА ЭКОНОМИКИ РЕСПУБЛИКИ БЕЛАРУСЬ

Т.С. Космыкова
Главный специалист ОАО «Банк БелВЭБ», заместитель декана инженерно-экономического факультета по научно-исследовательской работе студентов БГУИР, ассистент кафедры экономической информатики, магистр экономических наук, магистр технических наук

Белорусский государственный университет инфоматики и радиоэлектроники, Республика Беларусь
E-mail: t.kasmykova@gmail.com

Abstract. This article is about the bankruptcy risk of enterprises and the problems of solvency analysis of Belarusian organizations using the methodology of solvency research of enterprises in the country. The author made an attempt to reveal the specific analysis of bankruptcy risk for the organization of the Republic of Belarus using ratio analysis, and to indicate the strengths and weaknesses of this method. She tried to highlight the disadvantages of using a ratio analysis in Belarus, and proposed some measures to improve its conduction. As an alternative method of solvency research and identifying the risk of bankruptcy the author proposes analytical methods as the most prospective and focuses on econometric models of binary choice.

В настоящее время экономическая ситуация в Республике Беларусь характеризуется наличием рецессии, которая длится более, чем два года, сокращением реальных доходов населения, высоким уровнем внешней и внутренней долговой загрузки. В этой связи возрастает актуальность получения объективной информации об экономическом состоянии и степени устойчивости предприятий реального сектора экономики.

В настоящее время разработано большое количество математических моделей, способных оценивать риск банкротства. Многие из них эффективно используются в мировой практике, многие находятся в стадии уточнения, что способствует проведению их дальнейшего исследования.

Модели наступления неблагоприятного для организаций последствия можно условно разделить на пять групп: балансовые, рыночные, основанные на макроэкономических показателях, рейтинговые и гибридные [1].

В целях разработки модели прогнозирования риска банкротства предприятий Республики Беларусь была выбрана нелинейная логистическая регрессионная модель типа logit.

Нелинейная логистическая регрессионная модель типа logit является моделью, в которой переменная принимает только два различных значения, используемых при исследовании влияния тех или иных субъективных и объективных факторов на наличие либо отсутствие некоторого признака.

Если исследование затрагивает n субъектов, то есть если имеется n наблюдений, то факт наличия или отсутствия такого признака в i–м наблюдении удобно индексировать числами 1 (наличие признака) и 0 (отсутствие признака). Тем самым можно определить индикаторную (дихотомическую, бинарную) переменную y, которая принимает в i –м наблюдении значение y_i. При этом y_i=1 – при наличии рассматриваемого признака у i –го субъекта и y_i= 0 – при

отсутствии рассматриваемого признака у i –го субъекта [2 – 4].

Нелинейная логистическая регрессионная модель типа logit задана следующей формулой [5]:

$$G(y_{it}) = \frac{1}{1+e^{y_{it}}} \quad (1)$$

где G (y_i) – результирующий показатель модели (функция стандартного логистического распределения),

y_i – результирующий показатель линейной регрессионной модели (основания бинарной логистической регрессионной модели),

e – основание натуральных логарифмов, приблизительно равно значению 2,718281828.

Основанием нелинейной логистической регрессионной модели является линейная регрессионная модель со следующей спецификацией:

$$y_{it} = \beta_0 + \beta_1 * X_{1t} + \beta_2 * X_{2t} + \ldots + \beta_n * X_{nt} + \varepsilon_t, \quad (2)$$

где y_{it} – результирующий показатель линейной регрессионной модели,

β_0 – свободный член модели (константа),

β_1, β_2, β_n – веса при количественных и качественных показателях модели,

X_{1t} , X_{2t}, X_{nt} – факторы линейной регрессионной модели.

Факторами линейной регрессионной модели могут выступать количественные и качественные показатели.

Моделирование риска банкротства предприятий проводилось на основании бухгалтерской и иной отчетности за период с 2012 по 2016 годы.

Из имеющихся данных при моделировании использовались данные за 3 года (2012, 2013, 2014). Период для моделирования выбран исходя из стабильности работы экономики Республики Беларусь в данные периоды и отсутствие шоковых состояний в реальном секторе экономики. При этом выборка за 2014 будет обучающей для построения модели, а выборки за 2012 и 2013 годы – обучающие для тестирования построенной модели.

Для моделирования выбраны предприятия, относящиеся к сегменту средний и малый бизнес. Моделирование проводится с поправкой на специфику ведения бухгалтерского учета (УСН и бухгалтерский учет и отчетность на общих основаниях). Моделирования с поправкой на отраслевую принадлежность не производится.

Выборка формируется из предприятий Республики Беларусь, относящихся к малому и среднему бизнесу по историческим данным прошлых периодов. 30% организаций выборки относятся к клиентам, которые являются экономически несостоятельными (банкротами), оставшаяся часть клиентов относится к клиентам с отсутствием признака экономической несостоятельности (дефолта). Лизинговые компании и банки и кредитно-финансовые организации в выборке не участвуют.

В качестве зависимой переменной (эндогенной) в линейной регрессионной модели использована качественная переменная (бинарная), определяемая по историческим данным прошлых периодов. Под дефолтом будем понимать наличие невыполнения нормативных значений количественных показателей и срабатывание качественных показателей за период, а также наличие задолженности, просроченной свыше 90 дней. Данные ограничены клиентами отдельного банка Республики Беларусь. При этом значение 1 будет приниматься, если корпоративный заемщик не относится к дефолтным клиентам (не является банкротом), значение 0 – когда у корпоративного заемщика дефолтное состояние.

В качестве факторов модели прогнозирования риска банкротства изначально использованы 26 количественных и 10 качественные показателей, наилучшим образом описывающих финансовое состояние организаций.

Дальнейший анализ факторов позволил сократить их количество до 4 количественных и 1 качественного фактора.

К количественным факторам относятся:

–коэффициент обеспеченности собственными оборотными средствами, характеризующий наличие у субъекта хозяйствования собственных оборотных средств, необходимых для его финансовой устойчивости;

–коэффициент финансовой независимости (автономии), показывающий долю активов корпоративного заемщика, которые покрываются за счет собственного капитала (обеспечиваются собственными источниками формирования);

–коэффициент абсолютной ликвидности, показывающий долю краткосрочных долговых обязательств, которая может быть покрыта за счет денежных средств и их эквивалентов;

–темп прироста выручки от реализации продукции, товаров, работ, услуг (коэффициент), характеризующий изменение выручки от реализации продукции, товаров, работ, услуг за выбранный промежуток времени.

В случаях, когда расчет коэффициентов невозможен (например, деление на 0, отсутствуют данные для расчета и т. п.) либо в расчете отсутствует экономическая суть, значение по такому коэффициенту принимается равным 0 (нулю).

Коэффициент обеспеченности собственными оборотными средствами (К2) рассчитывается по следующей формуле:

$$K_2 = \frac{СК + ДО - ДА}{КА} \quad (3)$$

где СК – собственный капитал (значение строки 490 бухгалтерского баланса (далее – форма 1));

ДО – долгосрочные обязательства (значение строки 590 формы 1);

ДА – долгосрочные активы (значение строки 190 формы 1);

КА – краткосрочные активы (значение строки 290 формы 1).

Коэффициент финансовой независимости (автономии) (Кавт) рассчитывается по следующей формуле:

$$Кал = \frac{ДС}{КО} \quad (4)$$

где ДС – денежные средства и их эквиваленты (значение строки 270 формы 1);

КО – краткосрочные обязательства (значение строки 690 формы 1).

Темп прироста выручки от реализации продукции, товаров, работ, услуг (коэффициент) (Тпв) рассчитывается по следующей формуле:

$$Тпв = \frac{Выручка\ ТП - Выручка\ ПП}{Выручка\ ПП} \quad (5)$$

где Выручка ТП – выручка от реализации продукции, товаров, работ, услуг за текущий период (значение строки 010 графы 3 отчета о прибылях и убытках (далее – форма 2));

Выручка ПП – выручка от реализации продукции, товаров, работ, услуг за аналогичный период предыдущего года (значение строки 010 графы 4 формы 2).

Качественным показателем модели риска банкротства является кредитная история предприятия. Оценивается по данным за последние 24 месяца, анализируются все факты наруше-

ния кредитной дисциплины. Данный фактор может свидетельствовать о негативных изменениях в деятельности организации.

Кредитная история корпоративного заемщика (КИкз) характеризует качество исполнения обязательств перед банками и объем данных оценивается на основании анализа данных о выходе предприятия на просрочку по обязательствам перед банками.

Кредитная история корпоративного заемщика проставляется следующим образом: значение 0 (ноль) присваивается в случае, если кредитная история клиента оценивается как негативная, значение 1 (один) присваивается в случае отсутствия кредитной истории (признается нейтральной) либо кредитная история оценивается как удовлетворительная, значение 2 (два) присваивается, если кредитная история корпоративного заемщика признается положительной.

Непосредственное моделирование осуществлялось с помощью статистического пакета EViews 6.0, для чего были сгенерированы соответствующие временные ряды, затем выбран аппарат для моделирования – бинарные модели (вид: logit) [6].

Процесс оценивания сошелся после 7 итераций. Результаты представлены в таблицах 1 и 2.

Таблица 1. Результаты расчета неизвестных параметров модели, подлежащих оцениванию

Показатель	Коэффициент	Стандартная ошибка	z-статистика	Вероятность
константа	- 5,694692	1,093285	-5,208788	0,0000
коэффициент обеспеченности собственными оборотными средствами (k2)	5,664214	1,521267	3,723353	0,0002
коэффициент финансовой независимости (автономии) (kavt)	5,147904	1,446729	3,558306	0,0004
коэффициент абсолютной ликвидности (kal)	8,474396	4,786502	1,770478	0,0466
коэффициент, темп прироста выручки (tpv)	0,026168	0,007717	3,390934	0,0007
показатель кредитной истории клиента (f_kik)	1,912500	0,408178	4,685452	0,0000

Анализ данных паблицы 1 показал, что факторы, отраженные в модели, являются статистически значимыми, так как практически для всех показателей в модели выполняется условие: расчетное значение вероятности для z-статистики больше критического 0,05.

Таблица 2. Прочие результаты, характеризующие качество построенной модели

Критерии	Значения
R^2 Макфаддена	0,755084
Критерий Акаике	0,413490
Критерий Шварца	0,506342
LR-статистика	187,8073
Вероятность LR-статистики	0,00000
Критерий Ханна-Куина	0,450990
Логарифмическая функция правдоподобия	-0,179348
Ограниченная логарифмическая функция правдоподобия	-133,1808

Приведенные выше показатели также свидетельствуют о неплохом качестве модели. А коэффициент детерминации Макфаддена показывает, что модель способна описать выборку на 75,5%. Это свидетельствует о том, что подобранные факторы модели хорошо подогнаны под исходные данные.

Таким образом, после оценки модель приобрела следующий вид:

$$\hat{y}=-5,69+5,66x_1+5,15x_2+8,47x_3+0,026x_4+1,91x_5. \quad (6)$$

После подстановки обозначений, символизирующих факторы:

$$defolt=-5,69+5,66k_2+5,15kavt+8,47kal+0,026tpv+1,91f\_kik. \quad (7)$$

Для улучшения качества построенной модели была выполнена нормализация данных обучающей выборки по формуле:

$$X_i = \frac{x_i-\mu}{\sigma} \quad (8)$$

где X_i – нормализованное значение фактора,

x_i – нормализуемое значение фактора,

μ – среднеарифметическое значение факторов по выборке,

σ – стандартное отклонение распределения значений факторов по выборке.

Процесс оценивания нормализованных значений сошелся после 8 итераций.

Результаты перестроенной модели по нормализованным значениям показателей приведены в таблицах 3-4:

Таблица 3. Результаты расчета неизвестных параметров модели, подлежащих оцениванию

Показатель	Коэффициент	Стандартная ошибка	z-статистика	Вероятность
константа	7,875282	2,263110	3,479849	0,0005
коэффициент обеспеченности собственными оборотными средствами (k2)	5,780472	2,437049	2,371915	0,0177
коэффициент финансовой независимости (автономии) (kavt)	8,746507	2,836842	3,083185	0,0020
коэффициент абсолютной ликвидности (kal)	60,75886	17,88267	3,397639	0,0007
коэффициент, темп прироста выручки (tpv)	17,95857	6,089267	2,949216	0,0032
показатель кредитной истории клиента (f_kik)	3,512899	1,305433	2,690984	0,0071

Анализ приведенных данных показал, что факторы, отраженные в модели, являются статистически значимыми, так как практически для всех показателей в модели выполняется условие: расчетное значение вероятности для z-статистики больше критического 0,05.

Таблица 4. Прочие результаты, характеризующие качество построенной модели

Критерии	Значения
R^2 Макфаддена	0,895837
Критерий Акаике	0,181484
Критерий Шварца	0,274335
LR-статистика	238,6166
Вероятность LR-статистики	0,000000
Критерий Ханна-Куина	0,218984
Логарифмическая функция правдоподобия	-13,87251
Ограниченная логарифмическая функция правдоподобия	-133,1808

Приведенные выше показатели также свидетельствуют об улучшении качества модели и увеличении ее прогностической способности, а коэффициент детерминации Макфаддена показывает, что модель способна описать выборку на 89,6%.

Таким образом, после оценки модель приобрела следующий вид:

$$\hat{y}=7,88+5,78x_1+8,75x_2+60,75x_3+17,96x_4+3,51x_5. \quad (9)$$

После подстановки обозначений, символизирующих факторы:

$$defolt=7,88+5,78k_2+8,75kavt+60,75kal+17,96tpv+3,51f\_kik. \quad (10)$$

Таким образом, пятифакторная линейная регрессионная модель, наилучшим образом описывающая обучающую выборку можно представить в следующем виде:

$$y_i = \beta_0+\beta_1*К2+\beta_2*Кавт+\beta_3*Кал + \beta_4*Тпв +\beta_5*КИкз, \quad (11)$$

где y_i – результирующий показатель линейной регрессионной модели,

β_0 – свободный член модели (константа),

β_1, β_2, β_3, β_4, β_5 – веса при количественных и качественных показателях модели.

Количественные значения свободного члена и весов при количественных и качественных показателях линейной регрессионной модели приведены в таблице 5.

Таблица 5. Количественные значения свободного члена и весов при факторах модели

Название показателя	Обозначение в модели	Количественное значение
Свободный член	β0	7,875282
Вес для показателя К2	β1	5,780472
Вес для показателя Кавт	β2	8,746507
Вес для показателя Кал	β3	60,75886
Вес для показателя Тпв	β4	17,95857
Вес для показателя КИкз	β5	3,512899

Модель прогнозирования риска банкротства рассчитывается для каждого предприятия. Сначала рассчитывается основание модели (пятифакторная линейная регрессионная модель) на основе данных бухгалтерской отчетности, представленной предприятием, и прочих данных.

Полученный интегральный показатель линейной регрессионной модели по отдельному предприятию в дальнейшем участвует в расчете результирующего показателя нелинейной регрессионной логистической модели типа logit, который может принимать значения в диапазоне [0; 1].

Литература

[1]. Космыкова, Т.С. Проблемы моделирования риска банкротства предприятий / Т.С. Космыкова // Материалы IX Международной научно–практической конференции молодых исследователей «Содружество наук – 2013», 23–24 мая 2013 / г. Барановичи, Республика Беларусь – 2013. – С. 37 – 39.

[2]. Космыкова, Т.С. Методы оценки риска банкротства предприятий // Наука и инновации. 2015, №2. С. 42 – 46.

[3]. Космыкова, Т.С. Выбор оптимального метода для выявления банкротства предприятий // Наука и инновации. 2015, №3. С. 42 – 45.

[4]. Алёхина, А.Э., Космыкова, Т.С. Моделирование риска банкротства с использованием моделей бинарного выбора / А.Э. Алёхина, Т.С. Космыкова // Сборник статей VIII Международной научно–практической конференции «Наука – промышленности и сервису», 7 – 9 ноября 2013 / г. Тольятти, Российская Федерация – 2013. – С. 199 – 208.

[5]. Космыкова, Т.С. Моделирование риска банкротства при помощи моделей бинарного выбора / Т.С. Космыкова // Материалы XXIII Международной научно–практической конференции «Управление в социальных и экономических системах», 15 мая 2014 / г. Минск, Республика Беларусь – 2014. – С. 139 – 141.

[6]. Носко, В. П. Эконометрика для начинающих (дополнительные главы) / В. П. Носко. – Москва: ИЭПП, 2005. – 379 с.

ПРЕДОБРАБОТКА БОЛЬШИХ ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ МЕТОДА ОЦЕНКИ ЧЕЛОВЕКА В УСЛОВИЯХ РИСКА

А.Л. Раднёнок
Ассистент кафедры инженерной психологии и эргономики БГУИР, магистр технических наук, аспирант

В.С. Осипович
Доцент кафедры инженерной психологии и эргономики БГУИР, кандидат технических наук, доцент

И.Г. Шупейко
Доцент кафедры инженерной психологии и эргономики, кандидат психологических наук, доцент

К.Д. Яшин
Заведующий кафедой инженерной психологии и эргономики БГУИР, кандидат технических наук, доцент

Белорусский государственный университет информатики и радиоэлектроники, Республика Беларусь
E-mail: Raelag@tut.ru

Abstract. Software tool provides the processing capability to expert data collected using human risk assessment method.

Целью работы является разработка программного средства для предварительной обработки экспериментальных данных, полученных в результате компьютеризированной методики оценки человека в условиях опасности, для дальнейшей оценки.

В качестве входных данных используется набор файлов с расширением *.csv, отдельный файл соответствует одному испытуемому. Содержимое файла имеет определенный формат, показанный в таблице 1.

Таблица 1. Формат содержимого файлов данных испытуемых

№ строки	Содержание				
1	Тип стимула (линия/дуга)				
2	Контактные данные испытуемого				
3	№ п/п	Время, с	Расстояние, мм	Наличие ошибки	Отклонение, мм
…	…	…	…	…	…

В качестве выходных необходимо получить средние значения данных по столбцам и количество ошибок для каждого испытуемого в виде списка для разных видов стимулов (линия/дуга), также построить график отклонения от целевой позиции представленной на мониторе в зависимости от номера стимула для отдельных испытуемых.

Для достижения цели было разработано программное средство обработки экспериментальных данных метода оценки человека в условиях риска, основанного на методе исследования реакции на движущийся объект. Программное средство реализовано на языке программирования C# в среде разработки Microsoft Visual Studio 2015.

Программное средство позволяет решать ряд следующих задач:

– загрузка файлов испытуемых;

– формирование списка испытуемых по средним значениям показателей для стимула «линия»;

– формирование списка испытуемых по средним значениям показателей для стимула «дуга»;

– формирование списка испытуемых по каждому стимулу;

– просмотр экспериментальных данных отдельного испытуемого;

– построение графика отклонения в зависимости от номера стимула;

– сохранение обработанных данных в Excel-файле;

Для загрузки списка испытуемых необходимо выбрать пункт меню «Файл → Открыть папку», выбрать из списка директорию, содержащую в себе файлы с экспериментальными данными. Результат выполнения изображён на рисунке 1.

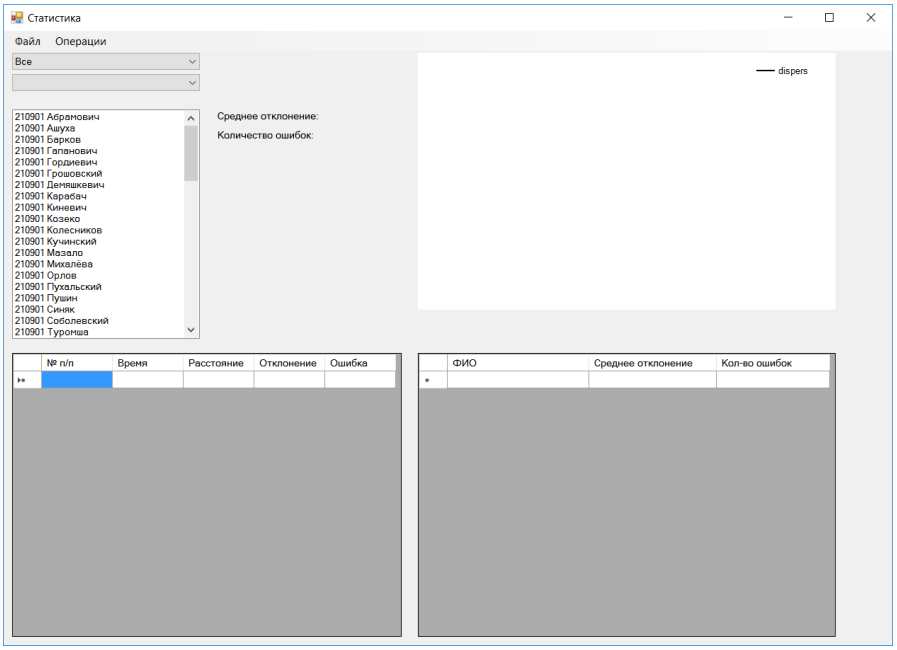


Рис. 1. Главное окно приложения

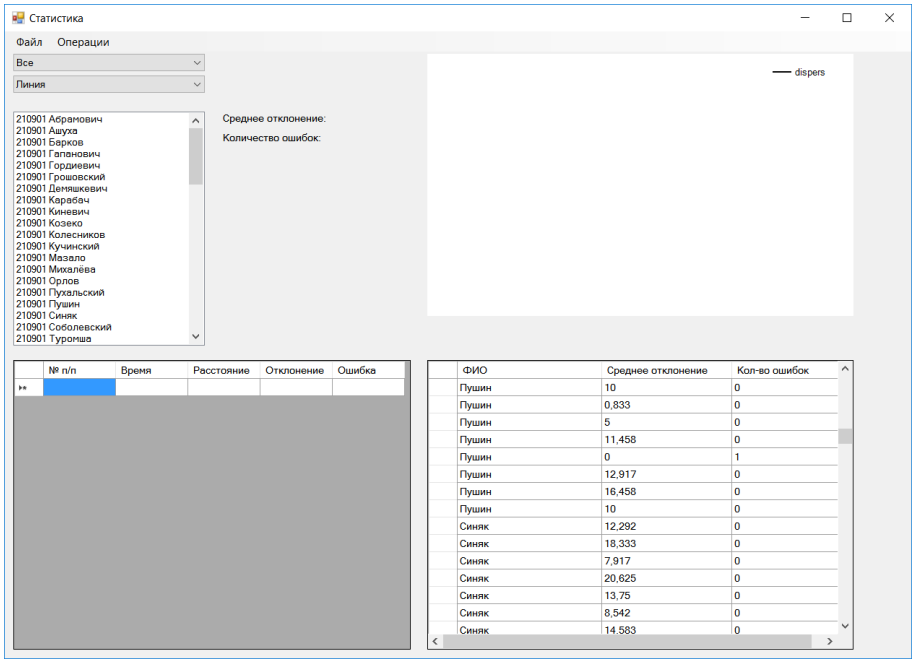


Рис. 2. Список испытуемых по отдельным стимулам вида «Линия»

Для формирования списка испытуемых по отдельным стимулам необходимо в выпадающем списке выбрать вид стимула, после чего сформируется список в зависимости от вида стимула: «Линия» или «Дуга». Результат формирования списка для стимула «Линия» представлен на рисунке 2.

Просмотр данных для отдельного испытуемого осуществляется выбором его из списка, приложение предоставит нужные данные в виде таблицы, и построит график зависимости отклонений от номера стимула.

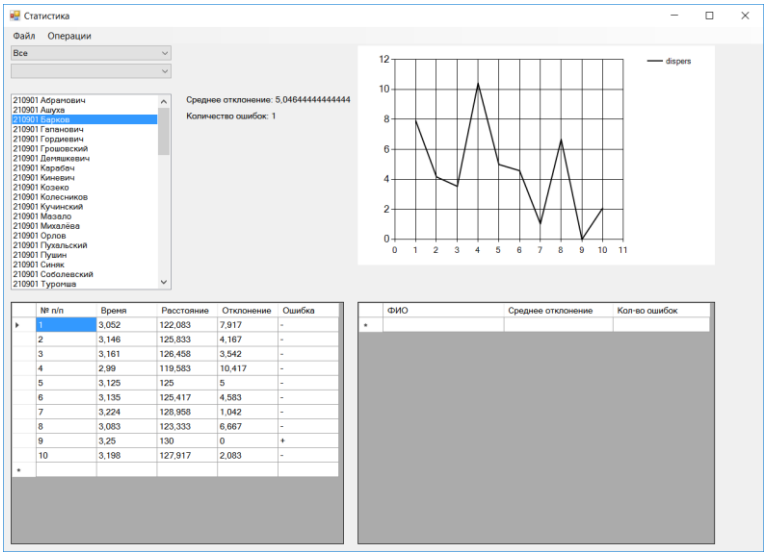


Рис. 3. Просмотр экспериментальных данных отдельного испытуемого

Для предоставления статистической информации по всем испытуемым для отдельного вида стимула необходимо выбрать пункт меню «Операции → Результаты по линии» или «Операции → Результаты по дуге». На рисунке 4 и рисунке 5 и представлены статистические данные по стимулу «Линия» и «Дуга» соответственно.

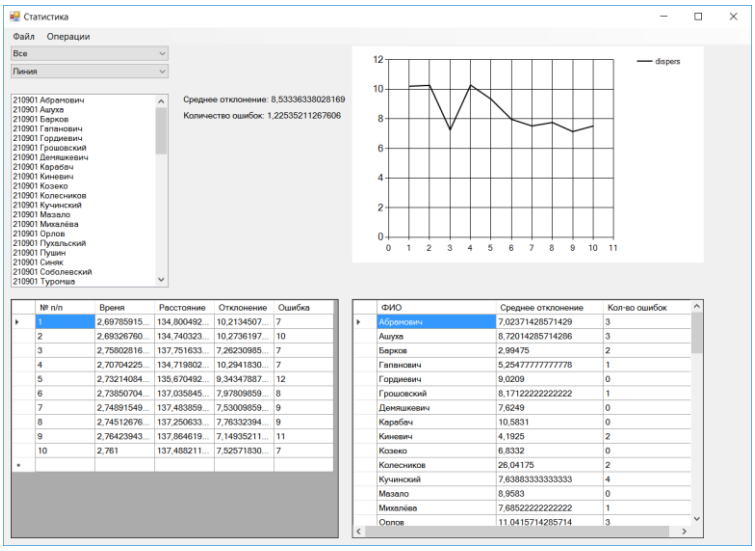


Рис. 4. Статистическая информация по испытуемым для стимула «Линия»

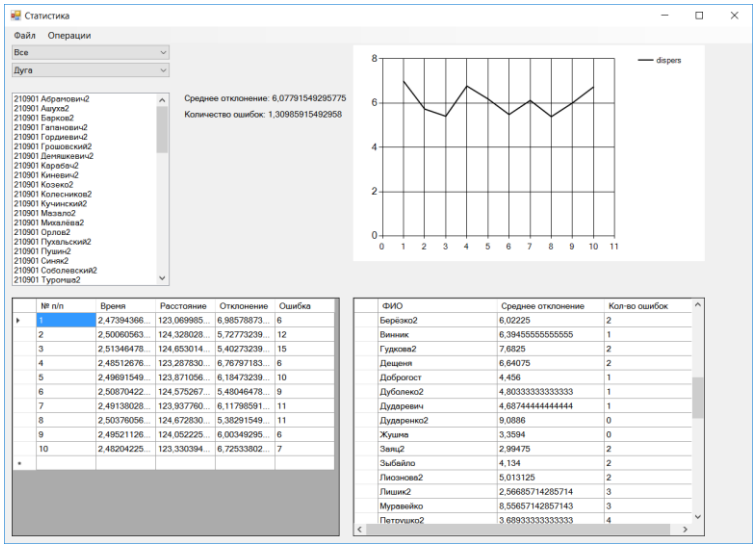


Рис. 5. Статистическая информация об испытуемых для стимула «Дуга»

Также есть возможность сохранить данные из таблиц в виде Excel-файла, для этого необходимо выбрать пункт меню «Файл → Сохранить…». Помимо этого все данные из таблиц можно скопировать, для этого необходимо выделить нужные данные из таблиц и выбрать пункт меню «Операции → Копировать»

В результате работы разработано программное средство для предварительной обработки данных, полученных в ходе эксперимента, посвящённого исследованию психофизиологических характеристик человека в условиях риска. Программное средство реализовано на языке программирования C# в среде разработки Microsoft Visual Studio 2015.

ИСПОЛЬЗОВАНИЕ BEACONS ДЛЯ ПОСТРОЕНИЯ СИСТЕМЫ НАВИГАЦИИ ВНУТРИ ЗДАНИЙ

Л. А. Лось
Магистрантка БГУИР

Н.А. Волорова
Заведующая кафедой информатики БГУИР, кандидат технических наук, доцент

Белорусский государственный университет информатики и радиоэлектроники, Республика Беларусь
СООО «Интетикс Бел», Республика Беларусь
E-mail: lyubov.los@gmail.com, volorova@bsuir.by

Abstract. The aim of the work was to review the possible approaches for the implementation of navigation and positioning systems in buildings and through this, establish the most optimal solution in all respects. As a result, a beacon-based implementation was chosen. The general scheme of work was considered and the main technologies for implementing the system were selected. The difficulties of development and methods for their solution were determined.

С развитием архитектуры новые строения становятся более обширными и нередко имеют сложную структуру. Ориентирование в таких сооружениях для человека, впервые попавшего в них, нелегкая задача. Системы навигации внутри зданий стремятся облегчить ориентирование на новом пространстве, дать дополнительную информацию, построить маршрут и точки следования. Для производителей и продавцов товаров такие системы становятся новыми горизонтами для рекламы их товаров. Данные собранные такими системами могут использоваться в маркетинге для анализа посещений магазинов, наплывов покупателей, пользующиеся успехом товары. Системы позиционирования используются в таких местах, как музеи, крупные торговые и развлекательные центры, выставки, туристические объекты и т.д.

Таким образом, существует множество причин для использования систем внутреннего позиционирования, но при этом существуют проблемы поиска решений вместо систем спутникового позиционирования, балансировки стоимости, сложности реализации и других факторов.

Задачей исследования стало рассмотрение и анализ возможных подходов к решению и реализация системы навигации в здании.

Так, для реализации таких систем могут использоваться различные варианты. Для определения местоположения повсеместно используются спутниковые системы навигации, однако они имеют критические недостатки, не позволяющие при определенных условиях доходить сигналу до приемника, поэтому они непригодны для определения положения внутри зданий.

Инерциальные системы основаны на модели движения человека. Если мы знаем откуда начинается движение, куда и как быстро движется объект, то можно рассчитать, где он окажется через некоторое время. Системы основываются на данных, полученных с помощью гироскопов и акселерометров смартфона. Однако для таких систем необходимо знать начальную точку, и со временем накапливающуюся погрешность приходится сверять с другими источниками.

Системы, основанные на измерении магнитного поля с помощью компаса смартфона, требуют предварительной калибровки в помещении и могут подвергаться влиянию металлов и магнитов.

Использование Wi-Fi точек не дает достаточной точности, погрешность может составлять до 25 метров. Конфигурирование же такой сети, которая позволит достичь хорошего уровня точности, потребует значительных материальных затрат.

Системы, основанные на использовании Beacon-маячков, предоставляют достаточную точность при приемлемом уровне затрат. Типичный Beacon-маячок имеет достаточно компактные размеры и способен проработать от одной батарейки в течение нескольких лет. Дальность действия зависит от конкретной модели и настроек и в среднем составляет от 10 до 40 метров. Цена одного такого устройства обычно не превышает 30$.

Система навигации на основе Beacons обычно строится по следующей схеме. По всей территории помещения устанавливаются Bluetooth-маяки, по заранее известным координатам в пространстве. Пользовательское приложение получает информационные сообщения от этих маяков через установленный промежуток времени. Исходя из полученных данных и мощности полученного сигнала, циклично определяется текущее положение принимающего устройства.

Периодичность вещания данных от маяков может быть настраиваемым параметром, обычно используются значения от 100 мс и реже. Периодичность выдачи данных влияет на продолжительность работы устройства, и, конечно же, на точность определения местоположения. Чем чаще клиентское приложение получает данные, тем точнее строится маршрут его следования. С другой стороны, увеличение частоты вещания Beacon (100мс – выдача данных 10 раз в секунду), значительно увеличивает поток данных, которые нужно обрабатывать в режиме реального времени. Таким образом, одно клиентское приложение принимает ежесекундно данные некоторого количества маяков, в радиус действия которых попадает, передает эти данные на сервер, где они должны быть сохранены для дальнейшей обработки, проанализированы в данный момент времени, рассчитаны координаты и отданы обратно на клиентское приложение. Количество хранимых данных на сервере постоянно увеличивается, их хранение позволяет проводить аналитические исследования за определенные периоды времени, позволяя получать данные по наиболее популярным маршрутам (в торговом центре, например), общему количеству посетителей в конкретном месте в конкретное время и т.д.

Поскольку нигде точно не определено, какое точное количество данных уже подпадает под понятие «Big Data», под этим термином понимается не только большие объемы данных, но и набор технологий для их сбора, обработки и хранения. Есть мнение, что такие технологии должны решать следующие проблемы:

–Уметь обрабатывать больше данных, по сравнению со стандартными сценариями (данные, генерируемые физическими датчиками);

–Уметь работать с быстро поступающими данными в больших объемах (постоянно увеличивающимися в количестве);

–Уметь работать с плохо структурированными данными (подразумевается, что алгоритмы могут получать на вход не всегда структурированную информацию – для определения координат сигналы с разного количества маяков).

Исходя из таких положений, можно сказать что построение систем навигации относится к категории задач в понятиях «Big Data».

Рассмотрим общую схему работы системы, изображенную на рисунке 1.

Рис. 1. Общая схема системы

Для реализации системы необходимы непосредственно маяки (или устройства, имитирующие их работу). Под Bluetooth Low Energy Beacon будем понимать миниатюрные батарейные устройства, работающие на основе BLE, для передачи небольшого объема статической или динамической информации. Такие устройства часто предназначены для беспрерывной работы в течение нескольких лет, которая обеспечивается технологией BLE, подразумевающей низкое энергопотребление, достигаемое сокращением времени передачи данных и погружением устройства в режим сна между передачей пакетов [3].

Рассматривая доступные варианты работы устройств BLE, можно выделить режимы, основанные на соединении с другими устройствами (connection-based) и режимы с однонаправленной передачей или приемом, не требующие соединения [1]. Для реализации маячка видятся два режима:

–Периферийное устройство (Peripheral (slave)) – устройство, периодически отправляющее информационное сообщение и принимающее входящие соединения. После активации соединения «следует» за центральным устройством и регулярно обменивается с ним данными.

–Широковещательный передатчик (Broadcaster) – устройство, не подключаясь, периодически отправляет пакеты любому желающему их получить.

В обоих случаях пакет вещаемых данных содержит одинаковую информацию, за исключением одного флага, показывающего устройство соединяемое или нет. В поставленной задаче было бы логичней использовать не соединяемые маячки, которые просто передают информацию. Такой вариант позволяет использовать устройство в режиме минимального потребления энергии, при этом главная функция маяка для нашей системы будет выполняться – трансляция пакета данных с некоторой частотой. Данные рассылки не будут изменяться после начальной установки, что позволяет однозначно идентифицировать каждый маяк.

Передаваемые данные имеют формат, определенный спецификацией Bluetooth. В Таблице 1 показана значимая для нас часть данных пакета, транслируемого Beacon.

Такой набор данных, состоящих из идентификатора группы маяков, мажора и минора, позволяет точно определить конкретный Beacon и, с помощью мощности маяка, расстояние до него.

Поскольку Beacon – это BLE устройство, его может заменить на любое другое устройство с установленным BLE-чипом и программным обеспечением, реализующим функции маяка. Таким образом, приложение на смартфоне, поддерживающем стандарт Bluetooth 4.0 LE

(Low Energy) [2], является альтернативной заменой маяка. Чем мы и воспользуемся. Для программной реализации Beacon-маяка будем использовать Xamarin для одновременного создания эмулятора Beacon на несколько платформ.

Таблица 1. Данные, транслируемые маяком.

Данные	Размер	Значение
Преамбула	4 байта	Префикс пакета, сообщающий, что это именно Beacon
Идентификатор группы маяков (UUID)	16 байт	Идентификатор, позволяющий отличать, например, маяки одного магазина от второго. Т.е. все маяки, расположенные в торговом зале одного магазина, будут иметь одинаковый идентификатор
Мажор	2 байта	Уникальный идентификатор подгруппы маяков в рамках UUID, позволяет выделять в группу маяки, находящиеся в одном зале большого магазина
Минор	2 байта	Идентификатор, позволяющий определить конкретный маяк
Эталонное значение мощности маяка (TX Power)	2 байта	Сила сигнала на расстоянии в 1 метр от маячка, используется для определения расстояния до пользователя

В качестве клиентского приложения чаще всего выступает мобильное приложение, имеющее определенный функционал. Во-первых, возможность сканирования входящих информационных пакетов от маячков. Это подразумевает, что мобильное устройство поддерживает Bluetooth 4.0, а само приложение реализует функцию наблюдателя (Observer) – сканирует эфир в поисках объявлений от вещателей, но не инициализирует соединение при этом.

Во-вторых, передача полученных от маяков данных на сервер, для вычисления по этим данных текущих координат. И, наконец, прием рассчитанных координат и отображение текущего местоположения.

Для разработки клиентского приложения будем использовать платформу Xamarin, позволяющую вести разработку сразу для нескольких платформ. Приложение будет содержать несколько проектов. Один из них «core», в котором будет содержаться основная логика приложения (общение с сервером, модели, протоколы и т.д.), и отдельные UI проекты для каждой платформы. Клиентское приложение будет отправлять данные от маяков на сервер, где будут вычисляться текущие координаты. Для серверной части выберем решения и сервисы, предоставляемые бесплатной облачной платформой IBM Bluemix.

При вычислении координат могут возникать трудности и при использовании системы на основе Beacon. Системы, основанные на BLE, полагаются на электромагнитные волны, что может приводить к получению неверных данных с дальних маяков. Для улучшения позиционирования прибегают к некоторым уловкам. Например, использование одновременно нескольких подходов для уменьшения количества ошибок. Таким образом, вместе с Beacon можно снимать данные с других датчиков мобильного устройства такими, как акселерометр, магнитометр.

Также используется специальная расстановка маяков, когда в каждый момент времени клиентское устройство расположено в зоне видимости определенного количества маяков. Для определения координат используются алгоритмы определения местоположения – подходы к решению задач определения местоположения на основе мощностей сигналов, посылаемых Beacon-маячками [4]. Алгоритмы, которые могут быть использованы:

–Ближайшая точка доступа», когда клиенту просто присваиваются координаты точки, излучающей наиболее мощный сигнал.

–«Центроид». Представляет собой вычисление геометрического центра плоской фигуры,

образованной несколькими маяками.

Координаты клиента определяются следующим образом:

$$\begin{cases} X_0 = \frac{1}{N}\sum_{i=1}^{N} X_i \\ Y_0 = \frac{1}{N}\sum_{i=1}^{N} Y_i \end{cases}, \quad (1)$$

где N – количество маяков,

X_i, Y_i – координаты маяков

–Латерация». Геометрический подход, основан на вычислении расстояний между искомой точкой и, как минимум, еще тремя точками (рисунок 2), с решением системы нелинейных уравнений. Для вычисления координат пользовательского устройства, необходимо решить систему уравнений:

$$r_i = \sqrt{(X_i - X_0)^2 + (Y_i - Y_0)^2} \quad (2)$$

где r_i – расстояние от клиентского устройства до маяков.

Для нахождения расстояний используется модель распространения радиоволн, требующая калибровки параметров, зависящих от особенностей среды:

$$PL(d) = P_t - P(d) = PL(d_0) + n10lg\frac{d}{d_0}, \quad (3)$$

где d – расстояние до клиентского устройства,

PL(d) – потеря мощности сигнала на расстоянии d.

P_t – мощность передатчика,

P(d) – мощность сигнала на приемнике на расстоянии d,

d_0 – расстояние 1 метр,

n – коэффициент распространения сигнала в среде.

Первые два алгоритма отличаются простотой реализации и требуют только знания местоположения маяков. Однако их недостатком является низкая точность, так как не учитывается мощность сигналов. Преимуществом алгоритма «Латерации» является достаточно высокая точность, при соответствующих параметрах среды. Недостатком же является необходимость построения более тщательной модели.

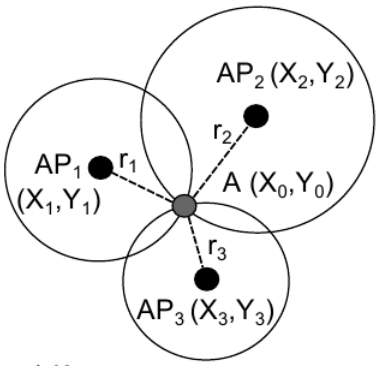


Рис. 2. Круговая латерация

В результате исследования были рассмотрены актуальность и необходимость построения систем навигации и позиционирования в помещениях, наиболее популярные подходы, позволяющие ее реализовать.

Системы, реализованные с использованием Beacon, являются конкурентоспособной альтернативой другим способам реализации систем навигации внутри помещений, но имеют свои плюсы и минусы.

В данной работе были рассмотрены различные технологии, с помощью которых достигается реализация системы

Литература

[1]. Kevin Townsend, Carles Cufí, Akiba, Robert Davidson, Getting Started with Bluetooth Low Energy, O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

[2]. Bluetooth Low Energy adopted specification [Электронный ресурс] – Режим доступа: – URL https://www.bluetooth.org/en-us/specification/adopted-specifications (дата обращения 12.04.2017).

[3]. Маячки Bluetooth Low Energy [Электронный ресурс] – Режим доступа: – URL http://www.compel.ru/lib/ne/2015/11/3-mayachki-bluetooth-low-energy.

[4]. Р.М. Миниахметов, А.А. Рогов, М.Л. Цымблер. Обзор алгоритмов локального позиционирования для мобильных устройств. Вестник Южно-Уральского государственного университета. Серия: Вычислительная математика и информатика. Выпуск № 2 / том 2 / 2013

МНОГОКРИТЕРИАЛЬНАЯ МАРШРУТИЗАЦИЯ ИНФОРМАЦИОННЫХ ПОТОКОВ

Н.И.Листопад
Заведующий кафедрой информационных радиотехнологий, доктор технических наук, профессор

А.В.Короткевич
Декан факультета радиотехники и электроники кандидат технических наук, доцент

С.Ю.Михневич
Доцент кафедры информационных радиотехнологий кандидат физико-математических наук, доцент

А.А.Хайдер
Стажер БГУИР

Белорусский государственный университет информатики и радиоэлектроники, Республика Беларусь
E-mail: seth22@yandex.ru

Absract. Traditionally, path selection within routing is formulated as a shortest path optimization problem. The objective function for optimization could be any one variety of parameters such as number of hops, delay, cost...etc. The problem of least cost delay constraint routing is studied in this paper since delay constraint is very common requirement of many multimedia applications and cost minimization captures the need to distribute the network. So an iterative algorithm is proposed in this paper to solve this problem. It is appeared from the results of applying this algorithm that it gave the optimal path (optimal solution) from among multiple feasible paths (feasible solutions

Маршрутизация информационных потоков формулируется как оптимизационная задача поиска кратчайшего пути. Целевая функция может быть любой из множества разнообразных параметров, таких как количество узлов, величины задержки, стоимости и др. [1]. Отдельной проблема маршрутизации является выбор оптимального пути при ограничениях по задержке и по стоимости, так как требования по минимизации задержки являются очень распространенными для многих мультимедийных приложений.

В работе [1] исследуется проблема поиска оптимальных маршрутов на графе мультисервисной телекоммуникационной сети. Для данных сетей, кроме полосы пропускания, должны приниматься во внимание такие параметры качества обслуживания (QoS), как потери пакетов, задержка пакетов, вариация времени задержки (джиттер). Задачу маршрутизации в мультисервисных сетях предлагается решать на основе критериев, учитывающих перечисленные параметры, согласно требованиям конкретных приложений.

Эта задача сформулирована как многокритериальная задача поиска маршрута с минимальной стоимостью, причем поиск выполняется только на подмножестве осуществимых путей, удовлетворяющих ограничениям на параметры качества сервиса. В работе предложена модификация алгоритма Дейкстры, которая позволяет осуществлять многокритериальный поиск оптимального маршрута с учетом ограничений на каждый критерий в отдельности, а также в случае, когда стоимость маршрута неаддитивна.

Важной проблемой, с которой столкнулись авторы статьи [1], это выбор весовых коэффициентов, помощью которых осуществляется свертка параметров, обеспечивающий заданные требования качества обслуживания (QoS), в комплексный коэффициент, в соответствии с которым и производится выбор оптимального пути.

В работе [2] представлен алгоритм поиска пути, для которого установлена минимальные

стоимость и задержка передачи информации (Delay-Constrained Least-Cost – DCLC – path). Т.е. рассматривается задача двухкритериальной маршрутизации, где в качестве оптимизационной функции выбраны два параметра: величина задержки в передаче информации и стоимость.

Рассмотрим данный вопрос более подробно. Пусть задана сеть в виде графа, у которой для каждой дуги, описывающей каналы передачи информации, определены величины задержки и стоимость. При этом два вышеназванных параметра свернуты в один с помощью единого комплексного весового коэффициента. Затем, используя данный коэффициент, применяется алгоритм Дейкстры для поиска кратчайшего пути.

Теоретически может быть доказано, что до тех пор, пока параметр выбран оптимальным образом, полученный кратчайший путь должен быть допустимым решением со стоимостью не больше, чем у пути с наименьшей задержкой (Least Delay - LD) [2]. На основании этого результата, используется эвристический алгоритм, который позволяет получить хорошие решения поиска кратчайшего пути на основании применения алгоритма Дейкстры. В целях дальнейшего повышения качества получаемого решения затем предлагаются два итерационных алгоритма, которые могут генерировать ряд параметров, постепенно улучшающих соответствующие решения.

1. Проблема поиска кратчайшего пути с наименьшей стоимостью. Любая сеть может быть представлена направленным граф G (V, E), где V есть множество узлов, и E есть множество каналов связи между ними. Предположим, что N = [V], и M =[E].

Вес *w* определяется как неотрицательное вещественное число w (e), описывающее каждый канал связи, т.е. W: E $\rightarrow R_0^+$. В частности, вес d: E $\rightarrow R_0^+$ называется задержкой, в то время как c: E $\rightarrow R_0^+$ называется стоимостью. Путь является конечная последовательность не повторяемых узлов p = (v_1,v_2,..., v_k) таких, что для $0 \le i < k$, существует связь от v_1 до v_{i+1}, т.е. $(v_i, v_{i+1}) \in E$. Канал $e \in p$ означает, что p проходит через канал связи e. Вес w, как задержка или стоимость, аддитивны, если вес пути p является суммой весов всех составляющих каналов связи вдоль этого пути,

$$w(p) = \sum_{e \in p} w(e) \quad (1)$$

В частности, задержка и стоимость пути p задаются двумя ниже представленными уравнениями:

$$d(p) = \sum_{e \in p} d(e) \quad (2)$$

$$c(p) = \sum_{e \in p} c(e) \quad (3)$$

В общем смысле, задержка по каналу связи есть среднее время передачи по этому каналу, в то время как стоимость может не взыматься при

передачи сообщения по этому каналу.

Приведем несколько определений [2].

Определение 1.

Заданы сеть *G (V, E),* источник $s \in V$ и узел назначения $t \in V$, заданы задержка и стоимость каждого канала связи, и ограничение по задержке – C_d .

Необходимо решить задачу поиск кратчайшего пути от *s* до *t* при минимальной стоимости с учетом следующих ограничений:

(i) $d(p) \le C_d$,

(ii) $c(p) \le C(q)$ для любого пути q от s до t, что удовлетворяет $d(p) \le C_d$,

(iii) Не существует пути q от s до t, для которого $c(p) = c(q)$,

В то время как $d(p) > d(q)$.

Следует отметить, что третье требование не является обязательным при решении задачи поиска оптимального пути при минимальной стоимости. Оно введено для того случая, когда возможно существования более одного решения для стандартной задачи. Для удобства, путь, который по крайней мере удовлетворяет первому требованию в приведенном выше определении, называется допустимым решением (или реальный путь); путь, который удовлетворяет всем трем требованиям, называется оптимальным решением (или оптимальный путь).

Следующее определения и условные обозначения необходимы для описания алгоритмов, которые будут предложены ниже.

Определение 2.

Даны два аддитивных веса w_1 и w_2, а также аддитивный вес $w = w_1(e) + \alpha w_2(e)$ означает, что для любого канала связи

$$w(e) = w_1(e) + \alpha w_2(e) \quad (4)$$

где $E \rightarrow R_0^+$.

Очевидно, что комплексный вес двух аддитивных весов также является аддитивным.

Определение 3:

Заданы узел источника информации *s* и узел назначения *t*, а также весовой коэффициент *w*. Это определяет функцию (или процедуру) Dijk (*w*), которая позволяет найти кратчайший путь *w* от *s* до *t* с помощью алгоритма Дейкстры. В частности, это эквивалентно следующему. Пусть на пути p_d= Dijk(*d*) задержка минимальная (LD путь), а путь p_c = Dijk(*c*) имеет минимальную стоимость (LC путь) между *s* и *t*. Нетрудно увидеть, что это соотношения $d(p_d) \leq d(p_c)$ и $c(p_d) \geq c(p_c)$ выполняются всегда.

Другая функция, которая будет использоваться в наших алгоритмах, это ModiDijk(*c,d*). Если существует несколько путей с различными задержками от *s* до *t*, функции ModiDijk(*c,d*) выберет тот из них, который имеет минимальную задержку. Это может быть сделано с помощью модифицированного алгоритма Дейкстры.

2. Идея единого комплексного весового коэффициента.

Основная идея предлагаемых алгоритмов состоит в решении задачи посредством объединения требования по задержке и стоимости посредством единого комплексного весового коэффициента и затем, используя алгоритм Дейкстры, в нахождении подходящего (кратчайшего) пути.

Рассмотрим проблему на простейших примерах рис.1, где необходимо решить кратчайший путь от *s* до *t* с величиной задержки, равной 8, и минимальной стоимостью – так называемый DCLC путь. Теперь, решая эту задачу вручную, требуется проверить все четыре пути между *s* и *t*. Легко определить, что LC путем является путь *s*-*u*-*t*, который имеет задержку 9 и таким образом является недопустимым. Путем LD является путь *s*-*v*-*t*, который имеет задержку 5 и стоимость 24. Хотя этот LD путь осуществим, он не является оптимальным решением, так как величина задержки не минимальна.

Введем комплексный весовой коэффициент вес $w = d + \alpha c$, который объединяет в себя задержку и стоимость. Вместо коэффициента *d*, определяющего величину задержки в передаче информации, могут быть использованы и весовые коэффициенты, определяющие другие параметры качества обслуживания, например, джжитер, полосу пропускания, вероятность потерь пакетов. Например, если будем рассматривать полосу пропускания и стоимость то, коэффициент $w = \omega + \alpha c,$ где $\omega -$ полоса пропускания. Аналогичные выражения можно записать и для других параметров, характеризующих качества обслуживания.

Покажем на примере комплексного коэффициента, объединяющего в себе задержку и стоимость, как это можно реализовать на практике.

Пусть $\alpha = 0.5$, то весовой коэффициент w будет ассоциироваться с путем, найденным с помощью алгоритма Дейкстры - *s-u-v-t*. Этот путь имеет задержку 8 и стоимость 16, и оказывается оптимальным.

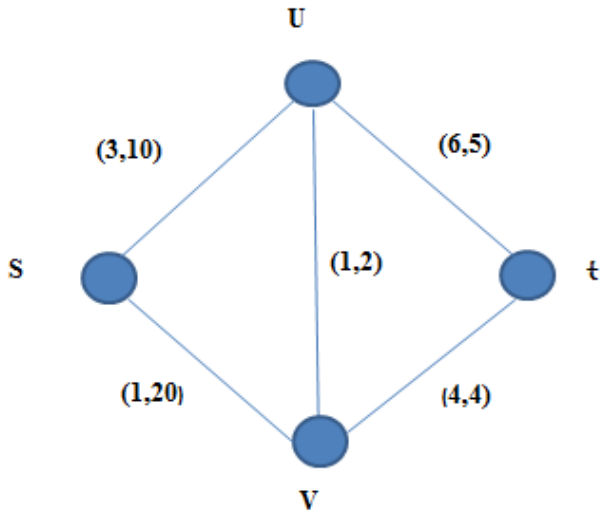


Рис.1. Иллюстрация проблемы поиска кротчайшего пути от s к t с минимальной стоимостью и минимальной задержкой, не превышающей 8

Этот пример показывает, что выбор соответствующего параметра α для построения комплексного весового коэффициента w, сводит DCLC задачу к задаче поиска кратчайшего пути, которая может быть легко решена с помощью алгоритма Дейкстры.

Ключевым вопросом для этой идеи является то, как выбрать параметра α для построения единого комплексного весового коэффициента w. Случайно выбранное значение α может привести к любым самым разнообразным решениям. Например, при α = 0,2 самый короткий путь w путь это *s-v-t*. В то время как при α = 2 кратчайшим путем становится путь *s-u-t*.

Заключение. Идея комплексного весового коэффициента была предложена для того, чтобы решать задачи QoS одноадресной маршрутизации. Данный подход может быть использован для разработки эвристических алгоритмов для задач поиска оптимального пути с минимальной задержкой, минимальной вариации задержки, обеспечением заданной полосы пропускания, минимальной вероятностью потерь и минимальной стоимостью передачи информации.

Литература

[1]. Н. И. Листопад, Ю. И. Воротницкий, А. А.Хайдер // Оптимальная маршрутизация в мультисервисных сетях телекоммуникаций на основе модифицированного алгоритма Дейкстры. // Вестник БГУ, серия 1. – 2015, № 1, с.70-76.

[2]. Waleed A. Mahmoud, Dheyaa J. Kadhim // A Proposal Algorithm to Solve Delay Constraint Least Cost Optimization Problem.// Journal of Engineering, University of Baghdad, V.19, № 1, January 2013. – P. 155-160.

СИСТЕМА ВЫСОКОТЕХНОЛОГИЧНОГО МАРКЕТИНГА НА ОСНОВЕ БОЛЬШИХ ДАННЫХ

В.В. Дершень
Студентка кафедры экономики БГУИР

В.А. Пархименко
Заведующий кафедрой экономики БГУИР, кандидат экономических наук, доцент

Белорусский государственный университет информатики и радиоэлектроники, Республика Беларусь
E-mail: sosna.victoria@gmail.com, parkhimenko@bsuir.by

Abstract. In the article, the system of technology intensive marketing based on big data, data mining, knowledge discovery and predictive analytics has been proposed.

В ходе непрерывного развития технологий маркетинговые процессы постоянно претерпевают изменения. Те из них, которые раньше выполняли маркетологи вручную, сегодня полностью или частично автоматизируются. Можно сказать, что маркетинг из сугубо управленческой (по сути, гуманитарной) науки, базирующейся во многом на интуиции и здравом смысле лица, принимающего решения, а также его компетенции в сфере психологии потребителей, получает существенный технический уклон и становится высокотехнологичной прикладной дисциплиной [1].

Согласно определению экономического словаря, высокие технологии – это технологии, развивающиеся в ходе НТР. Часто используют заимствованное из английского языка выражение хай-тек (high-tech), обозначающее процессы с использованием передовых технологий [2]. Высокие технологии требуют масштабного задействования научных и материально-технических ресурсов, представляют собой передовой рубеж развития науки и техники, воплощают в жизнь самые свежие открытия и изобретения [3].

К высоким технологиям можно отнести:

−беспроводные технологии;

−нанотехнологии;

−робототехнику;

−электронику;

−программное обеспечение, в частности исследования в области искусственного интеллекта;

−системы безопасности;

−навигационные технологии;

−экологически чистые технологии (альтернативные источники энергии и переработка отходов);

−социальные технологии (системы распространения новостей, когнитивистика);

−биотехнологии [4].

В научной литературе определение высокотехнологичного маркетинга до сих пор не было сформулировано. На основе полученной информации понятию высокотехнологичный

маркетинг можно дать следующее определение: высокотехнологичный маркетинг – это совокупность использующих высокие технологии методов продвижения и сбыта товаров и услуг.

Поскольку в научной литературе существуют разные подходы к определению маркетинга, то возможен еще один вариант трактовки понятия «высокотехнологичный маркетинг»: высокотехнологичный маркетинг – это совокупность процессов создания, продвижения и предоставления продукта или услуги покупателям, использующих высокие технологии, и управление взаимоотношениями с покупателями с выгодой для организации с использованием высоких технологий.

Следует отметить, что границы высокотехнологичного маркетинга постоянно смещаются. Это заложено в определении высоких технологий: то, что было ново и актуально несколько лет назад, сегодня может перейти в разряд обычных технологий.

С точки зрения авторов данной статьи, в настоящий момент новым в первую очередь выступают технологии, связанные с большими данными (big data) и интеллектуальным анализом данных (data mining, knowledge discovery, predictive analytics). Так, некоторые инструменты высокотехнологичного маркетинга прямо базируются на сборе и анализе больших данных, тогда как другие не имеют непосредственной связи с этими процессами, но могут быть связаны с Big Data. Например, QR-коды и NFC-метки могут приводить пользователей в места, где данные будут собраны.

Визуализация связи высокотехнологичного маркетинга с большими данными изображена на рисунке 1.

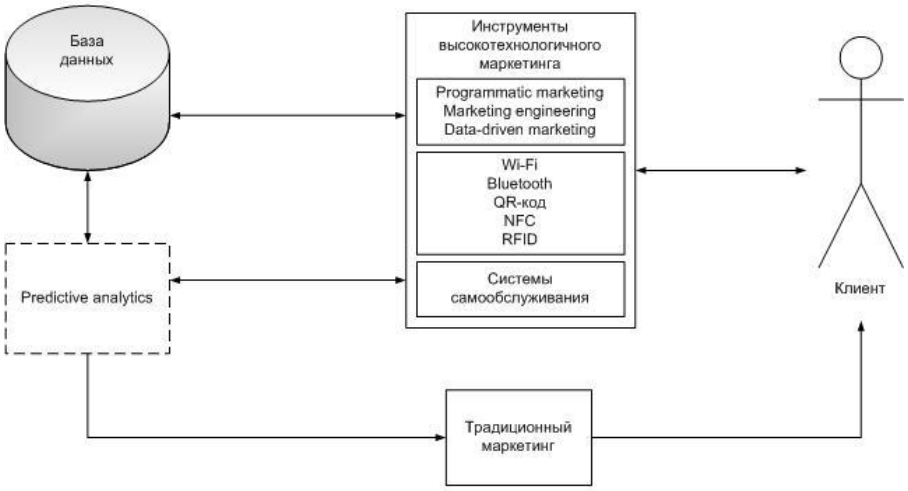


Рис. 1. Система высокотехнологичного маркетинга на основе больших данных

Инструменты высокотехнологичного маркетинга могут использоваться как в сборе информации о потребителях для дальнейшей работы с ней, так и для обратного воздействия по результатам анализа данных. Кроме того, собранные с помощью высокотехнологичных решений данные могут использоваться для воздействия на потребителей инструментами традиционного маркетинга.

Рассмотрим некоторые инструменты высокотехнологичного маркетинга, непосредственно связанные со сбором и анализом больших данных.

Маркетинг, основанный на данных. Data-driven маркетинг основывается на анализе массивов потребительских данных. В данном случае работа заключается в максимальной автоматизации и оптимизации внутренних и внешних процессов компании и отслеживании различ-

ных видов данных, таких как коэффициент оттока клиентов, уровень удовлетворенности клиентов, доля привлеченных потребителей, пожизненная ценность клиента, конверсия, прибыль, внутренняя норма доходности, окупаемость и т.д. Собранные данные позволяют прогнозировать, контролировать и управлять результатами компании [5].

Применение маркетинга, основанного на данных (data-driven стратегии) позволяет компании:

- точно определять целевую аудиторию рекламной кампании;
- максимизировать эффективность маркетингов вложений;
- повысить пожизненную ценность клиента;
- оперативно реагировать на изменения рынка.

Алгоритмический маркетинг. Алгоритмический маркетинг (programmatic marketing) – это автоматическое предложение цены на показы рекламы в режиме реального времени; совокупность методов закупки рекламы в интернете с использованием автоматизированных систем и алгоритмов для принятия решений о сделке без участия человека на основе социально-демографических и поведенческих данных о пользователях, имеющихся в распоряжении как площадки, так и рекламодателя [6]. Также в литературе можно встретить такие термины, как programmatic advertising, programmatic buying.

Принцип работы этой технологии следующий. Когда пользователь заходит на веб-страницу, ему демонстрируется реклама. За те доли секунды, пока идет загрузка сайта, система анализирует состав аудитории площадки, соотносит эти данные с таргетингом клиента, а также выбирает соответствующий рекламный формат. После запускается и проводит аукцион среди рекламодателей, целевой аудитории которых соответствует данный конкретный посетитель сайта и которые хотят показать ему свою рекламу. В ходе аукциона выбирается самая высокая ставка и дисконтируется до минимально необходимой для победы – это приводит к тому, что стоимость размещения рекламы, как правило, оказывается ниже, чем размер ставки. Затем реклама победителя загружается на сайт и демонстрируется пользователю.

Основным преимуществом данной технологии является значительное улучшение таргетинга, т.к. можно подобрать формат и содержимое в соответствии с ситуацией, в которой пользователь находится прямо сейчас [7].

Маркетинговая инженерия. Маркетинговая инженерия (marketing engineering) – системный подход к сбору данных и знаний для принятия эффективных маркетинговых решений с использованием технологий и моделей [8].

Основные принципы маркетинговой инженерии:

1 Маркетинговая инженерия направлена на решение проблем. Маркетологи часто сталкиваются с принятием решений относительно цены, упаковки или продвижения. В таких случаях маркетинговая инженерия покажет наилучший путь к решению проблемы.

2 Маркетинговая инженерия использует аналитическое программное обеспечение, которое позволяет выбрать наилучшую стратегию с помощью аналитического подхода.

3 Маркетинговая инженерия использует информационные технологии. Специально разработанное моделирующее программное обеспечение предлагает инструменты для принятия наилучших решений.

4 В маркетинговой инженерии решения принимаются на основе данных и знаний, что позволяет избежать субъективности и эмоциональных факторов при выборе стратегии [9].

Следующие инструменты и технологии имеют большое значение в системе высокотехнологичного маркетинга, так как помогают взаимодействовать с пользователями и, таким образом, связаны или могут быть связаны со сбором и анализом больших данных.

Wi-Fi-технологии в маркетинге. Wi-Fi (аббревиатура от английского Wireless Fidelity (беспроводная надежность) – это семейство протоколов беспроводной передачи данных IEEE 802.11 [10].

Принцип работы беспроводной сети построен на использовании радиоволн. Адаптер

беспроводной связи трансформирует информацию в радиосигнал и передает его в эфир через антенну. Беспроводной маршрутизатор принимает и делает обратное преобразование сигнала. Далее информация направляется в сеть Интернет по кабелю. Похожим образом осуществляется и прием информации. После получения информации из Интернета маршрутизатор преобразует ее в радиосигнал и отправляет через антенну на адаптер беспроводной связи устройства [11].

В маркетинге Wi-Fi можно использовать для рекламной коммуникации, сбора информации о клиентах, информировании о новинках, скидках, актуальных предложениях, программах лояльности, а также, благодаря собранным файлам cookie, поощрять дальнейшие визиты и активность.

Bluetooth-маркетинг. Bluetooth – производственная спецификация беспроводных персональных сетей, принцип действия которой основан на использовании радиоволн (Wireless personal area network, WPAN). Bluetooth обеспечивает обмен информацией между такими устройствами, как персональные компьютеры (настольные, карманные, ноутбуки), мобильные телефоны, принтеры, цифровые фотоаппараты, мышки, клавиатуры, джойстики, наушники, гарнитуры на надёжной, бесплатной, повсеместно доступной радиочастоте для ближней связи [12]. Bluetooth позволяет этим устройствам сообщаться, когда они находятся в радиусе до 10 метров друг от друга (дальность сильно зависит от преград и помех).

Bluetooth-маркетинг – способ реализации маркетинговых коммуникаций с использованием технологии Bluetooth в непосредственной близости от целевой аудитории [13]. Эта маркетинговая коммуникация возникла как реакция на массовое использование личных портативных устройств, которые поддерживают технологию бесконтактной передачи данных Bluetooth (мобильные телефоны, смартфоны, планшетные компьютеры).

Основными преимуществами этой технологии являются:

– ненавязчивость, запрос разрешения у каждого получателя для отправки контента, что формирует лояльность;

– бесплатная отправка, бесплатное получение информации;

– возможность формировать не только текстовые сообщения, но также использовать картинки, музыку, анимации, видеоролики, Java-приложения.

RFID-технологии в маркетинге. Радиочастотная идентификация (RFID) – это современная технология, используя которую информация необходимая для уникальной идентификации конкретного объекта, дистанционно записывается или считываются с наклеенной или встроенной в объект метки, с помощью радиоволн.

RFID-метка представляет собой миниатюрное запоминающее устройство. Она состоит из микрочипа, который хранит информацию, и антенны, с помощью которой метка эти данные передает и получает. Иногда RFID-метка имеет собственный источник питания (такие метки называют активными), но большинство меток его лишены (эти метки называют пассивными).

В памяти RFID-метки хранится уникальный номер и пользовательская информация. Когда метка попадает в зону регистрации, эта информация принимается считывателем, специальным прибором, способным читать и записывать информацию в метках [14].

NFC в маркетинге. NFC (Near Field Communication) – это технология беспроводной высокочастотной связи малого радиуса действия (до 10 см), позволяющая осуществлять бесконтактный обмен данными между устройствами, расположенными на небольших расстояниях: например, между считывающим терминалом и сотовым телефоном или пластиковой смарт-картой. Технология NFC базируется на RFID-технологии (Radio Frequency IDentification) [15].

Наиболее популярные варианты использования NFC технологии в мобильных телефонах:

1 Эмуляция карт – телефон прикидывается картой, например, пропуском или платежной картой.

2 Режим считывания – телефон считывает пассивную метку (tag), например, для интерактивной рекламы.

3 Режим P2P – два телефона связываются и обмениваются информацией.

В маркетинге технологии RFID и NFC применяются на рекламных мероприятиях для распространения промо-контента участниками мероприятия в социальных сетях [16]. Кроме того, возможно считывание меток телефоном для получения информации, участия в акциях, получения доступа к Wi-Fi.

QR-код в маркетинге. QR-код (quick response) – матричный код (двумерный штрих-код), разработанный и представленный японской компанией Denso-Wave в 1994 году [17].

QR-код определяется датчиком или камерой смартфона как двумерное изображение. Три квадрата в углах изображения и меньшие синхронизирующие квадратики по всему коду позволяют нормализовать размер изображения и его ориентацию, а также угол, под которым датчик расположен к поверхности изображения. Точки переводятся в двоичные числа с проверкой по контрольной сумме.

Основное достоинство QR-кода – это лёгкое распознавание сканирующим оборудованием, что дает возможность использования в торговле, производстве, логистике.

Возможностей применения QR-кодов в маркетинге очень много:

– получение дополнительной информации при сканировании кода;
– вовлечение клиентов для участия в розыгрышах и акциях;
– подарки за сканирование ссылок;
– оформление заказа или помещение товара в корзину при сканировании QR-кода;
– распространение контактной информации;
– получение обратной связи от потребителя;
– привлечение внимания клиентов с помощью вирусных кампаний, основанных на любопытстве.

Автоматизированные системы самообслуживания. Под системами самообслуживания понимаются как специальные сервисы на веб-сайтах [18], позволяющие клиентам самостоятельно получить необходимую информацию и выполнить некоторые операции, так и аппаратное обеспечение в офисах, магазинах, ресторанах, позволяющее клиентам делать заказы, совершать покупки и выполнять операции без помощи работников компании.

Примером онлайн-системы самообслуживания выступает Интернет-служба сервиса абонента (ИССА) компании velcom, где можно получить всю информацию о счете абонента и подключать или удалять услуги самостоятельно в своем аккаунте на сайте [19]. Офлайн-система самообслуживания применяется в филиалах Беларусбанк для обмена валют. Терминал предназначен для выполнения в режиме самообслуживания валютно-обменных операций с использованием наличных денежных средств, а также оперативного предоставления рекламно-справочной информации [20].

Использование таких систем позволяет оптимизировать многие процессы и ускорить обслуживание и повысить удовлетворенность клиентов.

Рекомендательные системы. Рекомендательные системы – программы (программные модули), которые на основе собранных данных пытаются предсказать, какие объекты или товары будут интересны клиенту.

Можно выделить два основных типа рекомендательных систем. В первом случае пользователю рекомендуются объекты, похожие на те, что этот пользователь уже употребил. Во втором случае для рекомендации также используются оценки и других пользователей, что часто дает лучший результат [21].

Грамотное использование рекомендательных систем в электронной коммерции позволяет увеличить среднюю выручку с пользователя и повысить удовлетворенность пользователей.

Онлайн-маркетинг (online-marketing, digital marketing, e-Marketing), мобильный маркетинг (mobile marketing) также в определенных случаях можно отнести к высокотехнологичному маркетингу.

Таким образом, высокотехнологичные решения в маркетинге в большинстве случаев либо уже связаны с Big Data, Data Mining и Predictive Analytics, либо потенциально могут быть, а с точки зрения авторов данной статьи, должны быть связанны с ними. Развитие в данном направлении позволит получать релевантную информацию о пользователях из многих источников, а анализ этой информации и грамотное ее применение сделают взаимодействие с потребителями еще более точным и индивидуальным, что позволит вывести маркетинг на новый, более высокий уровень.

Литература

[1]. Пархименко, В. А. Вызовы и пределы высокотехнологичного маркетинга / В. А. Пархименко // Веб-программирование и интернет-технологии. WebConf 2015 : материалы 3-й Междунар. науч.-практ. конф., Респ. Беларусь, Минск, 12–14 мая 2015 г. – Минск : БГУ, 2015. – С. 3–6.

[2]. Блэк, Дж. Экономика. Толковый словарь / Дж. Блэк. – М. : Инфра-М, 2000. – 848 с.

[3]. Высокие технологии – понятие и классификация [Электронный ресурс]. – Режим доступа : http://xbb.uz/Hi-Tech/Vysokie-tehnologii-ponjatie-i-klassifikacija/.

[4]. Высокие технологии – Википедия [Электронный ресурс]. – Режим доступа : https://ru.wikipedia.org/wiki/Высокие_технологии/.

[5]. Джеффри, М. Маркетинг, основанный на данных. 15 ключевых показателей, которые должен знать каждый / М. Джеффри. – М. : Манн, Иванов и Фербер, 2013. – 505 с.

[6]. What is Programmatic Marketing? [Электронный ресурс]. – Режим доступа : http://www.smartinsights.com/internet-advertising/internet-advertising-targeting/what-is-programmatic-marketing/.

[7]. Что такое программатик: модный термин или работающая технология? [Электронный ресурс]. – Режим доступа : http://digitalbee.com/blog/digital-marketing/programmatic-dlya-chaynikov-chto-takoe-programmatik-i-kak-on-rabotaet/.

[8]. Marketing engineering – Wikipedia [Электронный ресурс]. – Режим доступа : https://en.wikipedia.org/wiki/Marketing_engineering/.

[9]. Principles of Marketing Engineering [Электронный ресурс]. – Режим доступа : http://smallbusiness.chron.com/principles-marketing-engineering-77416.html/.

[10]. Пролетарский, А. В. Беспроводные сети Wi-Fi / А. В. Пролетарский, И. В. Баскаков, Д. Н. Чирков. – М. : БИНОМ, 2007. – 178 с.

[11]. Росс, Дж. Wi-Fi. Беспроводная сеть / Дж. Росс. – М. : НТ Пресс, 2007. – 178 с.

[12]. Bluetooth – Википедия [Электронный ресурс]. – Режим доступа : https://ru.wikipedia.org/wiki/Bluetooth/.

[13]. Bluetooth-маркетинг [Электронный ресурс]. – Режим доступа : http://www.e-xecutive.ru/wiki/index.php/Bluetooth-маркетинг/.

[14]. Власов, М. RFID. 1 технология - 1000 решений. Практические примеры использования RFID в различных областях / М. Власов. – М. : Альпина Паблишер, 2014. – 218 с.

[15]. Что такое NFC [Электронный ресурс]. – Режим доступа : https://faqhard.ru/base/17/01.php/.

[16]. Пархименко, В. RFID – новое слово в маркетинге?.. / В. Пархименко, М. Путилина // Маркетинг: идеи и технологии. – №1 (51), 2013. – С. 9–14.

[17]. QR-код – Википедия [Электронный ресурс]. – Режим доступа : https://ru.wikipedia.org/wiki/QR-код/.

[18]. Котлер, Ф. Маркетинг менеджмент. Экспресс-курс. 3-е изд. / Ф. Котлер, К. Л. Келлер. – СПб. : Питер, 2014. – 480 с.

[19]. ИССА. velcom [Электронный ресурс]. – Режим доступа : https://my.velcom.by/.

[20]. Терминал валютно-обменный Automated Currency Exchange Machine – IBA Беларусь [Электронный ресурс]. – Режим доступа : http://iba.by/products/banksystems/ATM/.

[21]. Как работают рекомендательные системы. Лекция в Яндексе [Электронный ресурс]. – Режим доступа : https://habrahabr.ru/company/yandex/blog/241455/.

К ВОПРОСУ О ПОДГОТОВКЕ ДАННЫХ ДЛЯ РЕШЕНИЯ ЗАДАЧ DATA MINING

Е.Н. Живицкая
Проректор по учебной работе БГУИР, кандидат технических наук, доцент

А.Т. Кусаинова[1]
Докторант Евразийского национальный университет имени Л.Н. Гумилева, магистр технических наук

В.А. Пархименко
Заведующий кафедрой экономики БГУИР, кандидат экономических наук, доцент

М.М. Татур
Профессор кафедры электронных вычислительных машин БГУИР, доктор технических наук, профессор

Белорусский государственный университет информатики и радиоэлектроники, Республика Беларусь
[1]Евразийский национальный университет им. Л.Н. Гумилева, Республика Казахстан
Email: jivitskaya@bsuir.by, ainurkussainova89@gmail.com, parkhimenko@bsuir.by, tatur@bsuir.by

Abstract. In the article, the problem of data preparation for data mining process has been described. Usually this stage of data preprocessing is skipped contrary to theoretical recommendations. Much more attention usually is paid to sophisticated methods and algorithms. This could lead to outcomes without any applicable meaning. In the article, authors have offered several ideas on dealing with main problems within data preparation step.

Введение. В учебной литературе по интеллектуальному анализу данных, как правило, излагаются классические формальные алгоритмы, а в научной литературе – их модификации и глубокое исследование различных аспектов их применения. При этом вопросам подготовки данных несправедливо уделяется мало внимания. Однако несложно показать, что подготовка (качество подготовки) данных может оказывать значительно более кардинальное влияние на конечный результат, нежели выбранный метод или модификация алгоритма.

Если рассмотреть (рис. 1) уже ставшее хрестоматийным изложение процесса Data Mining в рамках методологии CRISP-DM (Cross Industry Standard Process for Data Mining), то можно отметить, что в теории этап непосредственного применения конкретных алгоритмов Data Mining (на рисунке отмечено как «Modelling») является лишь одним из 6 этапов и носит соподчиненный характер по отношению к пониманию прикладных проблем из предметной области («Business understanding») и реализации принятых по итогам анализа решений («Deployment»).

В то же время, как показывает опыт, на практике в рамках многих проектов по Data Mining акцент переносится, напротив, сугубо на использование конкретных алгоритмов, а другие этапы (в том числе важный этап подготовки данных – Data preparation) опускаются.

В настоящей работе авторы затрагивают некоторые важные вопросы именно этого этапа, в частности проблему подготовки логических признаков и процедуру взвешивания и нормализации данных.

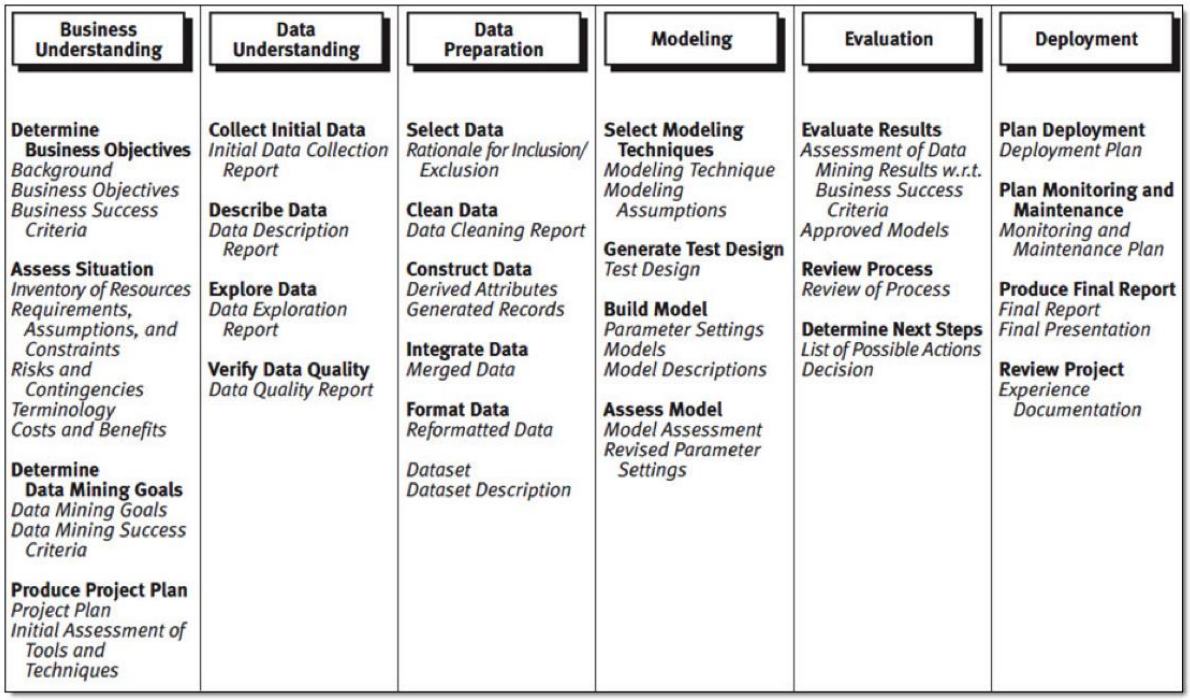


Рис. 1. Этапы, задачи и результаты Data Mining в соответствии с методологией CRISP-DM [1]

1. О совместном использовании логических и количественных признаков. Общепринято в анализируемых данных выделять количественные, логические, категориальные, порядковые и некоторые другие типы. Однако при рассмотрении конкретных алгоритмов Data Mining (например, кластеризации k-средних, классификации k-ближайших соседей и т.п.) речь идет исключительно о количественных данных. Как быть, если в задачах присутствуют данные различных типов [2], в частности логические?

Для ответа на этот вопрос рассмотрим формальный пример. Пусть имеется 3 объекта (образа), каждый из которых представлен двумя информативными признаками, при этом оба признака логические (признак либо присутствует, либо отсутствует у конкретного объекта).

Таблица 1 – Исходные данные

x_1	x_2	№ образа
0	0	–
0	1	1
1	0	3
1	1	2

$$O_1 = \overline{x_1}\, x_2$$
$$O_2 = x_1\, x_2$$
$$O_3 = x_1\, \overline{x_2}$$

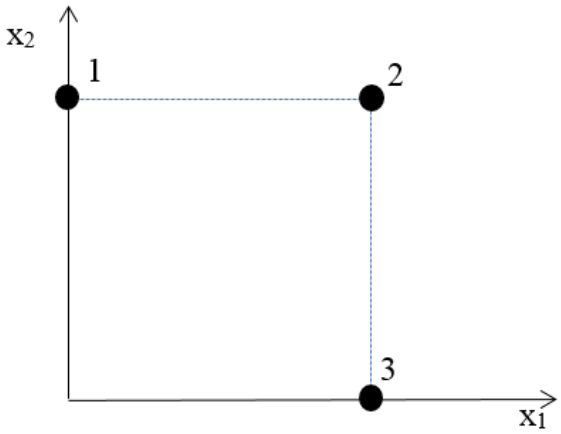


Рис. 2. Распределение образов с «чистыми» логическими признаками

Исходные данные по всем трем образам приведены в табл. 1. При их геометрической

интерпретации (рис.2) образы окажутся расположенными в вершинах квадрата (в общем случае, при наличии большого числа признаков – в вершинах гиперкуба). Максимальное число различимых образов – 2n, где n – число информативных признаков или размерность признакового пространства.

В такой постановке положение каждого из образов может быть описано булевым выражением, а кодовое расстояние между ними, при необходимости, может быть выражено в метрике Хэмминга.

Например, в коробке имеются детали, различающиеся по размеру – x_1 (большие и малые) и форме – x_2 (круглые и продолговатые). Необходимо записать правило принятия решения для автомата-сортировщика.

В данном случае признаки с точки зрения принятия решений являются «взаимно нейтральными»: большие круглые детали не лучше, не важнее, не значимее, чем, например, малые и продолговатые. Поэтому придание указанным признакам весовых значений лишено смысла, а булевы описания образов O_i и будут являться правилами принятия решений, чисто логических решений.

Но как только мы устанавливаем между этими признаками некоторое количественное соотношение, влияющее на принятие решения, тогда возникает подтип логических признаков, для которых правомерно оценивать геометрическое расстояние между образами.

Например, на ферме разводят кроликов. Имеющиеся кролики различаются по цвету (белый, серый) – x_1 и полу – x_2. Необходимо их сравнить в плане коммерческой (селекционной или др.) выгоды. Пусть признак x_1 при принятии решений в три раза важнее, чем признак x_2, тогда их взаимное расположение в пространстве (в частности, на плоскости) будет выглядеть, как представлено на рис.3. А значит, мы имеем все основания, чтобы вычислить евклидовы расстояния между образами и решать задачи кластеризации, ранжирования и др. известными алгоритмами.

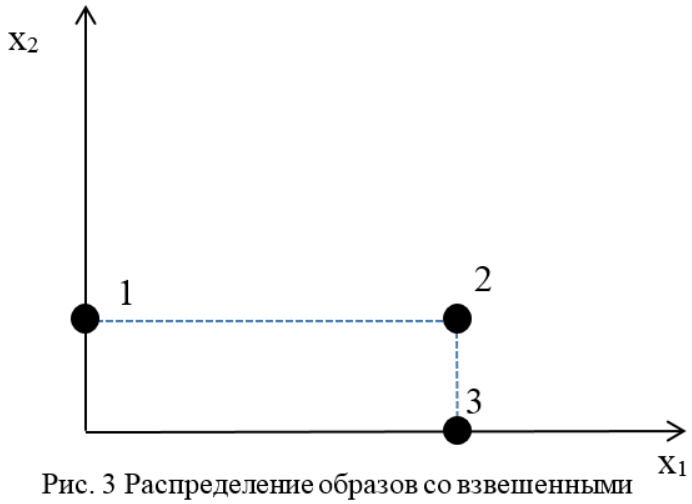


Рис. 3 Распределение образов со взвешенными логическими признаками

Продолжим пример. Пусть один из признаков (x_1) будет количественным, и мы к имеющимся трем добавим пятый и шестой образы, с параметрами, указанными в табл. 2. Так как признак x_1 остается в соответствии с условием важнее в три раза признака x_2, то распределение признаков графически можно будет представить таким, как показано на рис.4.

Если поменять соотношение весов признаков на противоположное (признак x1 станет менее весомым, чем признак x2 в три раза), то получим распределение, как показано на рис.5.

Таблица 2 – Дополнительные образы

x_1	x_2	№ образа
0	0	–
0	1	1
1	0	3
1	1	2
0,3	1	4
0,5	0	5

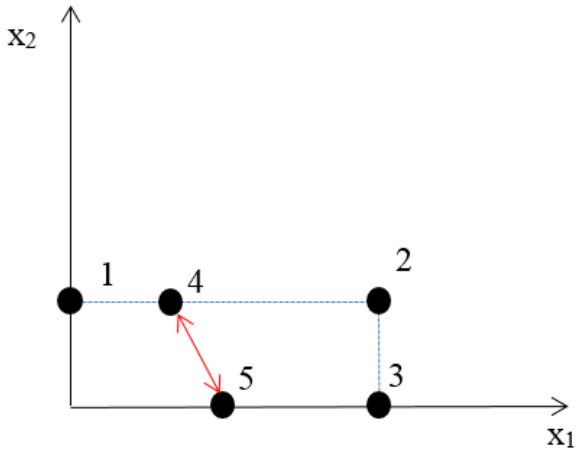


Рис. 4. Распределение образов с количественным и взвешенным логическим признаками, в соотношении 1:3

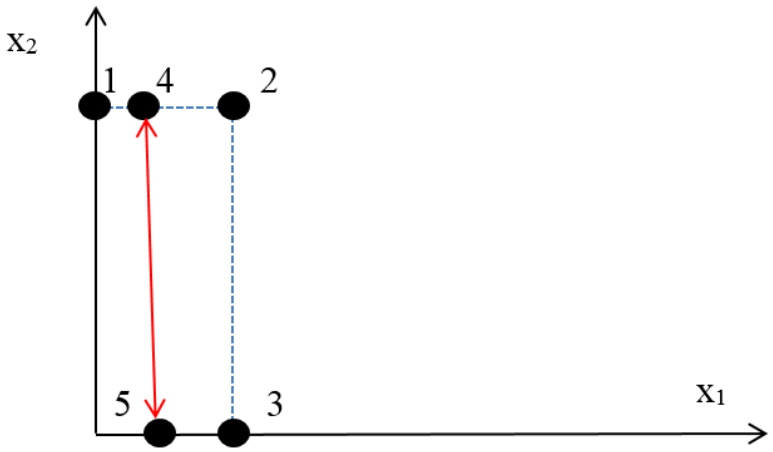


Рис.5. Распределение образов с количественным и взвешенным логическим признаками, в соотношении 3:1

Очевидно, что соотношения евклидовых расстояния между образами на рис. 4 и рис. 5 диаметрально противоположные:

$$\rho E\,(4,5) < \rho E\,(3,5),\ \rho E\,(4,5) < \rho E\,(2,4);$$

$$\rho E\,(4,5) > \rho E\,(3,5),\ \rho E\,(4,5) > \rho E\,(2,4),$$

а, следовательно, при решении задач Data Mining будут получены противоположные результаты.

2. Методические рекомендации по взвешиванию и нормализации информативных признаков. Давать строгие рецепты по взвешиванию и нормализации данных в виде методик и алгоритмов, по всей видимости, не возможно и даже не желательно [3]. Дело в том, что часть этапов подготовки данных, по своей сути, вообще не поддается формализации, а оставшаяся изобилует различными оговорками и частными случаями.

Поэтому заявленные «методические рекомендации» по взвешиванию и нормализации информативных признаков следует понимать, как некоторые эвристические правила, которые необходимо держать в поле зрения аналитику данных, особенно начинающему свою профессиональную карьеру.

Подобные правила авторы формулируют следующим образом:

1 Ранжировать признаки и установить веса означает, что необходимо сопоставить значимость всех используемых признаков в контексте конкретной решаемой задачи. Простейшим примером метода определения величины веса – является метод экспертных оценок. Однако не стоит забывать, что в некоторых случаях (при наличии априорной информации, репрезентативной выборки прецедентов и т.п.) значения весов могут быть «вычислены» в результате решения задачи Data Mining.

2 Чтобы унифицировать пользовательский интерфейс вычислительных систем при решении задач с различными диапазонами весов информативных признаков рекомендуется нормализовать область допустимых значений весов, т.е. вес максимально-значимого признака принять за единицу, а, минимально-значимого за ноль.

3 Изначально каждый из признаков представлен в оригинальной системе отсчета, исчисления, условных единицах и т.п. Для его использования в алгоритмах Data Mining в общем случае необходимо нормализовать одним из способов:

3.1 $x_i/(x_{max}-x_{min})$, т.е. текущее значение приводится к диапазону, вычисляемому из выборки;

3.2 $x_i/\Delta x$, т.е. текущее значение приводится к априори заданному диапазону Δx.

В зависимости от способа нормализации (3.1, 3.2) могут быть получены формально противоречивые результаты. Поэтому осуществлять нормализацию следует в контексте решаемой прикладной задачи.

Заключение. Логические признаки следует разделять на чисто логические и взвешиваемые, в зависимости от специфики решаемой задачи. Те и другие имеют фиксированное число кодовых комбинаций – 2^n, определяющее максимальное число распознаваемых образов.

Взвешиваемые логические признаки могут наряду с количественными информативными признаками использоваться для описания образов и без ограничений использоваться в алгоритмах Data Mining.

В ходе подготовки данных все признаки рекомендуется ранжировать и нормализовать как внутри собственной шкалы (области допустимых значений) так и глобально.

Некорректное соотнесение (взвешивание) информативных признаков, а также обработка ненормализованных данных может кардинально изменить результаты анализа, порой в большей степени, нежели выбор модификации примененного алгоритма Data Mining.

Вопросы геометрической интерпретации образов с категориальными признаками и их совместного использования с количественными станет предметом обсуждения в следующей работе. Еще одной важной проблемой на этапе подготовки данных является проблема выбора информативных признаков и связанная с ней проблема «очистки данных». В совокупности они напрямую влияют на результаты интеллектуального анализа и будут рассмотрены в дальнейших публикациях.

Литература

[1]. P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, R. Wirth. CRISP-DM 1.0: Step-by-step data mining guides [Электронный ресурс]. – Режим доступа: ftp://ftp.software.ibm.com/software/analytics/spss/support/Modeler/Documentation/14/UserManual/CRISP-DM.pdf

[2]. S. Ray. Simple Methods to deal with Categorical Variables in Predictive Modeling. [Электронный ресурс]. – Режим доступа: https://www.analyticsvidhya.com/blog/2015/11/easy-methods-deal-categorical-variables-predictive-modeling

[3]. L. A. Shalabi, Z. Shaaban, B. Kasasbeh. Data Mining: A Preprocessing Engine // Journal of Computer Science. – 2006. – №2. – P. 735-739.

АПРОБАЦИЯ РЕЗУЛЬТАТОВ НЕЛИНЕЙНОЙ РЕГРЕССИОННОЙ ЛОГИТ-МОДЕЛИ ПРОГНОЗИРОВАНИЯ РИСКА БАНКРОТСТВА ПРЕДПРИЯТИЙ И ОПРЕДЕЛЕНИЕ ЕЕ ОПТИМАЛЬНЫХ ПОРОГОВЫХ ЗНАЧЕНИЙ

Т.С. Космыкова
Главный специалист ОАО «Банк БелВЭБ», заместитель декана инженерно-экономического факультета по научно-исследовательской работе студентов БГУИР, ассистент кафедры экономической информатики, магистр экономических наук, магистр технических наук

Белорусский государственный университет инфоматики и радиоэлектроники, Республика Беларусь
E-mail: t.kasmykova@gmail.com

Abstract. This article is about the models of binary choice, that can be used to predict the risk of bankruptcy. There is some results of constructing models of binary choice in this article. This scientific material presents information about these models and their predictive ability, and also it includes the stages of model valuing. This article is focus on the critical values for the model for bankruptcy risk prediction and their determination. It is noted that the model is good for the bankruptcy risk prediction.

В качестве аппарата для моделирования риска банкротства предприятий реального сектора экономики целесообразно использовать нелинейную логистическую регрессионную модель.

В статье «Моделирование риска банкротства предприятий реального сектора экономики Республики Беларусь» [1] рассмотрен процесс построения данного типа моделей. Наилучший результат приобрела модель следующего вида:

$$\varphi(z) = \frac{e^z}{1+e^z} = \frac{1}{1+e^{-z}}, \quad (1)$$

где z – основание нелинейной логит-модели, представленное в виде пятифакторной регрессионной модели со следующей спецификацией:

$$z=7{,}88+5{,}78x_1+8{,}75x_2+60{,}75x_3+17{,}96x_4+3{,}51x_5, \quad (2)$$

где x_1 – коэффициент обеспеченности собственными оборотными средствами,
x_2 – коэффициент финансовой независимости (автономии),
x_3 – коэффициент абсолютной ликвидности,
x_4 – темп прироста выручки,
x_5 – показатель качества кредитной истории предприятия.

Прежде, чем приступать к определению пороговых значений, представляется целесообразным провести анализ распределения итоговых значений модели прогнозирования риска банкротства на обучающей выборке, на основании которой происходило ее построение.

Диапазон изменения результирующего показателя находится в пределах от 0 до 1. Чем

ближе расчетное значение к 0, тем ближе организация к состоянию банкротства, чем ближе расчетное значение к 1, тем ближе организация к состоянию финансовой устойчивости.

Обучающая выборка была сформирована из 219 предприятий: 65 относящихся к категории «банкрот» и 154 организаций без такого признака.

При этом, если значение результирующего показателя для i-ой организации было менее или равное 0,55, то значение классифицировалось как стремящееся к 0, а значение более 0,56 – как стремящееся к 1. Выявим ошибки 1-го и 2-го рода для обучающей выборки. Результаты содержатся в таблице 1.

Таблица 1. Результаты ошибок 1-го и 2-го рода для обучающей выборки

	Всего значений	Значение 1	Значение 0	Процент верно предсказанных значений
Значение 1	154	**149**	5	96,64%
Значение 0	65	5	**60**	92,31%
	219	Итого верных предсказаний: **209**	Итого неверных предсказаний: **10**	95,43%

Анализ данных таблицы показал, что на обучающей выборке модель дала высокие результаты и показала высокую прогностическую способность.

Проведем анализ качества предсказательной способности модели на выборках последующих периодов:

По состоянию на 01.01.2016 выявим ошибки 1-го и 2-го рода. Результаты приведены в таблице 2.

Таблица 2. Результаты ошибок 1-го и 2-го рода для выборки предприятий по состоянию на 01.01.2016

	Всего значений	Значение 1	Значение 0	Процент верно предсказанных значений
Значение 1	1556	**1519**	37	97,6%
Значение 0	81	5	**76**	93,8%
	1637	Итого верных предсказаний: **1595**	Итого неверных предсказаний: **42**	97,4%

На выборке по состоянию на 01.01.2016 модель дала высокие результаты и высокую прогностическую способность.

По состоянию на 01.04.2016 ошибки 1-го и 2-го рода следующие. Результаты приведены в таблице 3.

Таблица 3. Результаты ошибок 1-го и 2-го рода для выборки предприятий по состоянию на 01.04.2016

	Всего значений	Значение 1	Значение 0	Процент верно предсказанных значений
Значение 1	1317	**1173**	144	89,1%
Значение 0	104	8	**96**	92,3%
	1421	Итого верных предсказаний: **1269**	Итого неверных предсказаний: **152**	89,3%

На выборке по состоянию на 01.04.2016 модель также дала высокие результаты и хорошую прогностическую способность.

Снижение процента верно предсказанных показателей по сравнению с 01.01.2016 вызвано «выравниванием» показателей клиентов, стремящихся наилучшим образом «закрыть»

год.

По состоянию на 01.07.2016 выявим ошибки 1 и 2-го рода. Результаты приведены в таблице 4.

Таблица 4. Результаты ошибок 1-го и 2-го рода для выборки предприятий по состоянию на 01.07.2016

	Всего значений	Значение 1	Значение 0	Процент верно предсказанных значений
Значение 1	1248	**1139**	109	91,3%
Значение 0	98	9	**89**	90,8%
	1346	Итого верных предсказаний: **1228**	Итого неверных предсказаний: **118**	91,2%

На выборке по состоянию на 01.07.2016 модель дала хорошие результаты и хорошую прогностическую способность.

По состоянию на 01.10.2016 выявим ошибки 1-го и 2-го рода. Результаты приведены в таблице 5.

Таблица 5. Результаты ошибок 1-го и 2-го рода для выборки предприятий по состоянию на 01.10.2016

	Всего значений	Значение 1	Значение 0	Процент верно предсказанных значений
Значение 1	1171	**1097**	74	93,7%
Значение 0	101	9	**92**	91,1%
	1272	Итого верных предсказаний: **1189**	Итого неверных предсказаний: **83**	93,5%

На выборке по состоянию на 01.10.2016 модель дала хорошие результаты и высокую прогностическую способность.

По состоянию на 01.01.2017 выявим ошибки 1-го и 2-го рода. Результаты приведены в таблице 6.

Таблица 6. Результаты ошибок 1-го и 2-го рода для выборки предприятий по состоянию на 01.01.2017

	Всего значений	Значение 1	Значение 0	Процент верно предсказанных значений
Значение 1	918	**882**	36	96,1%
Значение 0	87	5	**82**	94,3%
	1005	Итого верных предсказаний: **964**	Итого неверных предсказаний: **41**	95,9%

На выборке по состоянию на 01.01.2017 модель дала высокие результаты и высокую прогностическую способность.

Увеличение процента верно предсказанных показателей по сравнению с 01.10.2016 вызвано «выравниванием» показателей клиентов, стремящихся наилучшим образом «закрыть» год.

Анализ полученных результатов показал, что модель в целом пригодна для прогнозирования риска банкротства предприятий.

Теперь представляется целесообразным произвести градирование итоговых значений интегрального показателя модели, расклассифицировав организации на большее количество групп, посредством установления пороговых значений для модели.

Для этого проанализируем обучающую выборку. Проведем графический анализ и построим диаграмму рассеивания значений, полученных в результате апробации модели на обучающей выборке (приведена на рисунке 1).

Из полученной диаграммы видно, что имеется возможность в зависимости от диапазона, в который попадает расчетное значение конкретного наблюдения, разделить полученные данные не менее, чем на 4 группы.

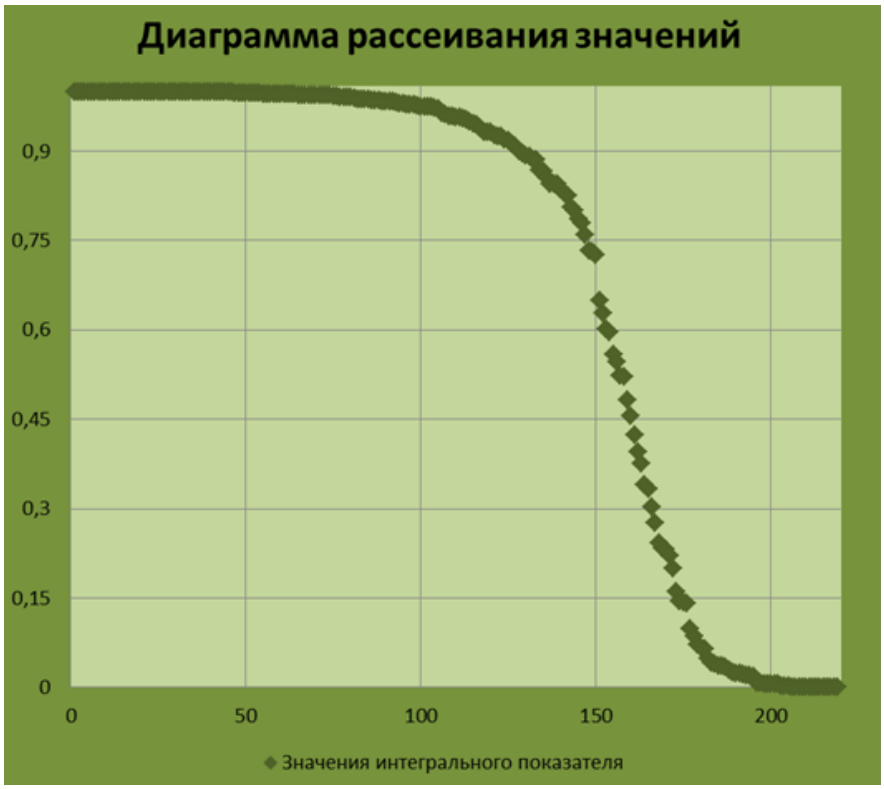


Рис. 1. Диаграмма рассеивания значений, полученных в результате апробации модели на обучающей выборке

Для более точного определения «оптимального» числа интервалов воспользуемся формулой Старджесса [2]:

$$k=\log_2 N+1=3{,}322\ \lg N+1, \quad (3)$$

где N – количество встречающихся в обучающей выборке повторяющихся значений результирующего показателя.

Исходя из полученного интегрального показателя модели, предприятия распределились следующим образом (приведено в таблице 7):

Таблица 7. Распределение предприятий в зависимости от значений интегрального показателя модели

Значение интегрального показателя	**0,00**	**0,04**	**0,07**	**0,08**	**0,13**	**0,15**	**0,19**
Количество предприятий	35	8	2	1	3	4	1
Значение интегрального показателя	**0,20**	**0,25**	**0,31**	**0,37**	**0,41**	**0,48**	**0,51**
Количество предприятий	1	1	1	1	1	1	1
Значение интегрального показателя	**0,53**	**0,54**	**0,57**	**0,69**	**0,86**	**1,00**	**Всего**
Количество предприятий	1	1	10	20	78	47	219

При N=20, значение k=5,29 (или 5 при округлении). Таким образом, для модели можно выделить 5 интервалов пороговых значений.

Определим шаг изменения интервалов по формуле:

$$h = \frac{n_{\max} - n_{\min}}{k}, \quad (4)$$

где n_{max} – максимальное значение результирующего показателя,

n_{max} – минимальное значение результирующего показателя,

k – оптимальное количество интервалов.

Следовательно, для анализируемых данных, шаг изменения интервалов h равен 0,19.

Таким образом, получаем следующие интервальные значения модели:

– от 0,00 до 0,19, в данный интервал входят организации категории «банкрот»,

– от 0,20 до 0,39, к данному интервалу относятся организации, близкие к банкротству

– от 0,40 до 0,59, в интервал входят предприятия, имеющие признаки финансовой неустойчивости,

– от 0,60 до 0,79, интервал стабильных организаций,

– от 0,80 до 1,00, интервал финансово устойчивых организаций.

При этом, для простоты отнесения предприятий к финансовой устойчивым предприятиям, предприятиям с признаками финансовой неустойчивости и предприятиям-банкротам представляется целесообразным указанные интервалы укрупнить следующим образом:

– от 0,00 до 0,39 – предприятия-банкроты,

– от 0,40 до 0,59 – предприятия, с признаками финансовой неустойчивости,

– от 0,60 до 1,00 –финансово устойчивые предприятия.

Литература

[1]. Космыкова Т.С. Моделирование риска банкротства предприятий реального сектора экономики Республики Беларусь / Т.С. Космыкова // Материалы XXIII Междунар. науч.-практ. конф. «BIG DATA and Advanced Analytics. Conference and EXPO», 3-4 мая 2017 / г. Минск, Республика Беларусь – 2017.

[2]. Выбор числа интервалов [Электронный ресурс], режим доступа: https://www.ami.nstu.ru/~headrd/seminar/xi_square/28.htm.– М., 2017.

СРАВНЕНИЕ РАЗЛИЧНЫХ ПОДХОДОВ К АНАЛИЗУ ТЕКСТА НА ПРИМЕРЕ ЗАДАЧИ ПРЕДСКАЗАНИЯ ОЦЕНКИ РЕСТОРАНА ПО ОТЗЫВУ ПОСЕТИТЕЛЯ

А.А. Шлеменков
Студент БГУИР

Белорусский государственный университет инфоматики и радиоэлектроники, Республика Беларусь
E-mail: alex.shlemenkov@gmail.com

Abstract. . In this research paper two common methods of texts analysis had been applied to the problem of restaurant review mark prediction. The results of the methods applied to the task had been analyzed and had been listed in table for comparison; reasons of different performance for specific pair of data and problem had been proposed.

Отслеживая тренды современного мира, можно заметить глубокий интерес к области искусственного интеллекта. Одной из областей в «машинном интеллекте» является естественная обработка языков (Natural Language Processing или NLP). Важно заметить, что область NLP является полезной не только для людей, которые тесно связаны с лингвистикой и языками, но и для бизнеса. Например, решив задачу поиска отрицательных отзывов на продукт, можно более оперативно реагировать на изменения.

В данной работе была рассмотрена задача, которая состоит в том, чтобы по тексту отзыва, который оставил посетитель, предсказать оставленную им оценку. Данные представляют собой текст отзыва и оценку в диапазоне от 1 до 5 (5 – лучший отзыв, 1 – худший) [1].

В ходе исследования были проанализированы два популярных метода анализа текстов. Один из них является «классическим» и основан на технике TF-IDF, другой был предложен относительно недавно и называется Word2Vec [2].

TF-IDF – статистическая мера, которая используется для оценки важности слова или сочетания из нескольких подряд идущих слов, которые, по сути, объединяются в одно уникальное. Выделение таких сочетаний часто очень полезно, так как позволяет «уловить» смысл отрицаний или устойчивых сочетаний, используемых в языке. Например, сочетание «не нравится» и слово «нравится» будут иметь абсолютно разный смысл в контексте документа, но если не учитывать такие фразы, то качество классификатора может сильно упасть. Мера TF-IDF каждого слова прямо пропорциональна количеству появлений в документе и обратно пропорциональна частоте появления слова во всех документах коллекции. Таким образом, чем чаще слово появляется в документе, тем выше его TF-IDF. И наоборот, если слово часто встречается во всех документах коллекции, например, общеупотребительная лексика, то даже с большим количеством появлений в документе значение TF-IDF этого слова будет мало. Данный подход позволяет отфильтровать часто встречающиеся элементы общеупотребительной лексики и, с другой стороны, выделить слова, которые встречаются редко во всем наборе, но часто в отдельных типах документов.

Для дополнительного сравнения был использован лемматизированный текст на входе. Общий смысл лемматизации заключается в приведении слов к начальной форме.

Способ анализа текстов под названием Word2Vec рассматривает проблему с другой стороны. В нем делается предположение о том, что слова, которые встречаются в схожих контекстах, имеют схожий смысл. Word2Vec каждому слову в коллекции ставит в соответствие вектор некоторой размерности (обычно это 100, 300, но размер вектора зависит от объема базы текстов, на которой был обучен Word2Vec). Этот вектор имеет смысл некоторой координаты в пространстве слов. Важное замечание состоит в том, что при изначальном предположении о контекстуальной близости слов получается, что вектора со схожим «смыслом» располагаются «рядом», а также существует возможность производить над такими векторами различные операции. Классическим примером является следующий: «король» - «мужчина» + «женщина» ~

«королева».

Оба метода были применены к данным. Для вычисления векторов слов был обучен Word2Vec на всех тренировочной базе отзывов, а после были усреднены вектора слов и обучена двуслойная нейронная сеть. В результате применения «классического» подхода использовался линейный классификатор с L1 регуляризацией. Результаты данного эксперимента в виде количества правильно классифицированных отзывов и матрицы ошибок представлены в таблице 1 и таблице 2:

Таблица 1. Матрицы ошибок

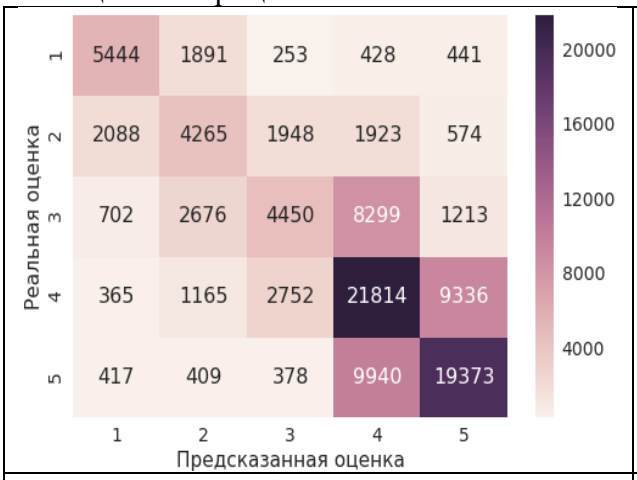
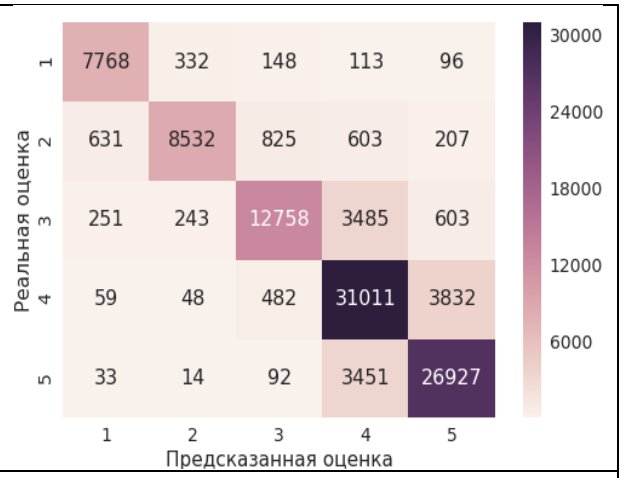
	
Word2Vec + NN	TF-IDF + LR

Таблица 2. Доля верно классифицированных отзывов

Используемый метод	Доля верно классифицированных отзывов, %
TF-IDF + LR	61,005
Lemmatization + TF-IDF + LR	58,23
Word2Vec + NN	54,852
Lemmatization + Word2Vec + NN	53,726

В таблице 1 в каждой из матриц ошибок на пересечении строки i (реальной оценки) и столбца j (предсказанной оценки) находится число, отражающее количество отзывов, которые были предсказаны классификатором как относящиеся к классу j, хотя на самом деле отзывы принадлежат классу i. Просуммировав все элементы на диагонали и поделив на сумму элементов в матрице можно получить долю правильно предсказанных оценок на тестовой выборке. Стоит отметить, что матрицы ошибок хорошо подходят для визуализации работы классификатора именно потому, что могут показать, где именно ошибается алгоритм.

Несмотря на новизну подхода, основанного на Word2Vec, как видно по результату, он работает не всегда хорошо. Объяснений этому может быть несколько: при усреднении векторов слов смысл их «размывается» слишком сильно. Это оставляет очень мало информации алгоритму для выделения каких-либо связей отзыва с оценкой. Также стоит отметить, что для каждого слова вычисляется его вектор, что не позволяет распознать «смысл» таких фраз, как, например, отрицания. Получается, что алгоритм не только не учитывает нужный «смысл» слов, но и учитывает его с обратным знаком. Подход, основанный на TF-IDF, сработал лучше из-за нескольких причин, самая важная которых: он учитывает сочетания слов. Следовательно, такие конструкции как «не нравится» распознаются как нечто отрицательное.

Заметим, что лемматизация несколько ухудшила результат. Это можно объяснить тем,

что при приведении слов к начальной форме теряется часть потенциально полезной информации.

Литература

[1]. Determine restaurant review sentiment [Электронный ресурс] – Режим доступа: https://in-class.kaggle.com/c/sentiment-analysis2

[2]. Distributed Representations of Words and Phrases and their Compositionality [Электронный ресурс] – Режим доступа: https://arxiv.org/pdf/1310.4546.pdf

СТАТИСТИЧЕСКИЙ АНАЛИЗ И МОДЕЛИРОВАНИЕ СТОИМОСТИ КВАРТИР НА ВТОРИЧНОМ РЫНКЕ ЖИЛОЙ НЕДВИЖИМОСТИ Г. МИНСКА

А.Э. Алёхина
Доцент кафедры экономической информатики БГУИР, кандидат экономических наук, доцент

Т.В. Федюкович
Ассистент кафедры экономической информатики БГУИР

Белорусский государственный университет информатики и радиоэлектроники, Республика Беларусь
E-mail: Ae.alekhina@gmail.com, Tatsiana.k@tut.by

Abstract. The purpose of this work is improvement of processes of economic and statistical modeling of residentials of real estate on the basis of multivariate statistical analysis. The statistical analysis of the residentials of real estate market in Minsk is carried out in the context of administrative districts of the city, types of buildings and number of rooms. Apartments were divided into clusters at a cost of one square meter. The regression model of the cost of residential real estate objects depending on the most significant indicators affecting the price of the apartment was constructed.

Рынок жилой недвижимости представляет собой сложную и разнородную сущность. Это тысячи квартир со своими уникальными свойствами, вплоть до вида из окон или уровня ремонта, состояния подъезда или наличия консьержа. Все эти квартиры расположены в разных концах города, каждый из которых наделен своей инфраструктурой и транспортной доступностью, имеет определенный уровень экологии и престижа.

В 2016 году в целом в г. Минске введено в эксплуатацию 753 тыс. кв. м. жилой недвижимости (первичный рынок). Это порядка 10,3 тысяч новых квартир, из которых 7,5 тысяч на счету коммерческих застройщиков [[2]]. Однако наиболее востребованным для населения является рынок вторичного жилья. Это оправдано следующими факторами: вторичный рынок изобилует предложениями, квартира уже физически и юридически готова к вселению новых хозяев сразу же после заключения договора купли-продажи, ремонт в обжитой ранее квартире не требует таких финансовых и временных затрат, как в случае с бетонной коробкой нового строения.

Рассмотрим структуру предложения квартир на вторичном рынке жилой недвижимости в зависимости от количества комнат (рисунок 1).

В предложении лидируют 2-комнатные квартиры. Их доля в общем числе составляет 34%. Доля предложений на однокомнатные квартиры составляет 32%. Всего на 6% меньше предложений по 3-комнатным квартирам. Значительно меньший выбор на вторичном рынке жилья предоставляется среди 4-комнатных и многокомнатных квартир. Их доля равна 7% и 1% соответственно.

В зависимости от местоположения, количество квартир, выставленных на продажу, изменяется следующим образом (таблица 1):

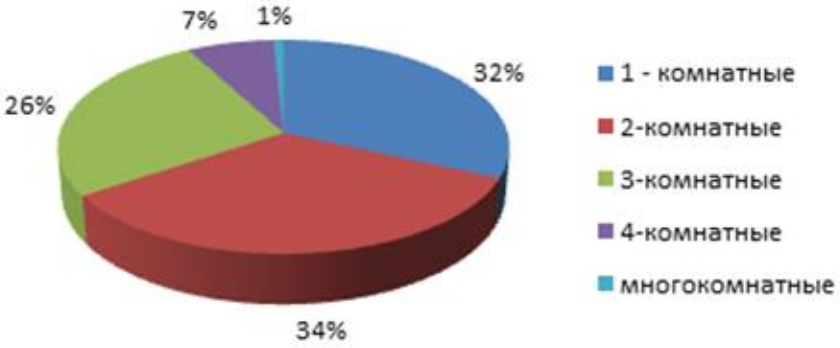


Рис. 1. Структура предложения квартир на вторичном рынке жилой недвижимости

Таблица 1. Распределение квартир по районам города на декабрь 2016 года

Район	*однокомнатные, шт.*	*двухкомнатные, шт.*	*трехкомнатные, шт.*	*четырехкомнатные, шт.*	*многокомнатные, шт.*
Заводской	152	184	132	45	1
Ленинский	132	215	175	49	5
Московский	246	264	217	74	7
Октябрьский	115	147	133	37	5
Партизанский	76	153	93	27	2
Первомайский	297	420	392	134	1
Советский	161	272	260	88	6
Фрунзенский	465	462	454	99	8
Центральный	207	326	288	99	12

Основная волна продаж в 2016 году пришлась на лето – количество регистрируемых сделок стало рекордным за последние десять лет. Больше всего сделок было заключено с квартирами в домах 1971–1999 года постройки, так как это самая распространенная группа из представленных на рынке.

За период с 7 февраля 2016 г. по 13 ноября 2016 г. на вторичном рынке жилой недвижимости г. Минска средняя рыночная стоимость 1 кв. м. общей площади квартир (на основе цены предложения к продаже) снизилась. В среднем по городу это снижение составило 76$ на 1 кв. м. общей площади квартир или, в относительном выражении, – 5,9%. Наибольшее снижение средних показателей наблюдается по трехкомнатным квартирам (на 7,2%), четырехкомнатным квартирам (на 6,8%). Темпы снижения рыночной стоимости 1 кв. м. общей площади в однокомнатных и двухкомнатных квартирах составили 4,4% и 5,5% соответственно (таблица 2) [2].

За первые три месяца 2016 года цены упали на 10%. Это относится непосредственно к цене предложения на вторичном рынке. Цены на жилье зависят от стоимости и объемов выдаваемых кредитов (в первую очередь льготных), и, естественно, от уровня заработных плат населения.

Рассмотрим цену сделки. Цена сделки значительно отличается от цены предложения. Сумма сделки в результате торга может быть на 5–10% ниже цены предложения. На рисунке 2 отображена цена сделки выставленных на продажу квартир в зависимости от количества комнат.

Таблица 2. Динамика средней цены 1 кв. м. общей площади квартир различного типа в г. Минске

Тип квартир по количеству комнат	*Средняя цена на 07.02.2016, USD*	*Средняя цена на 13.11.2016, USD*	*Абсолютный прирост (+), снижение (-), USD*	*Темп прироста (+), снижения (-), %*
1-комнатные	1327	1269	-58	-4,4
2-комнатные	1270	1200	-70	-5,5
3-комнатные	1266	1175	-91	-7,2
4-комнатные	1256	1171	-85	-6,8
В среднем по городу	1279	1203	-76	-5,9

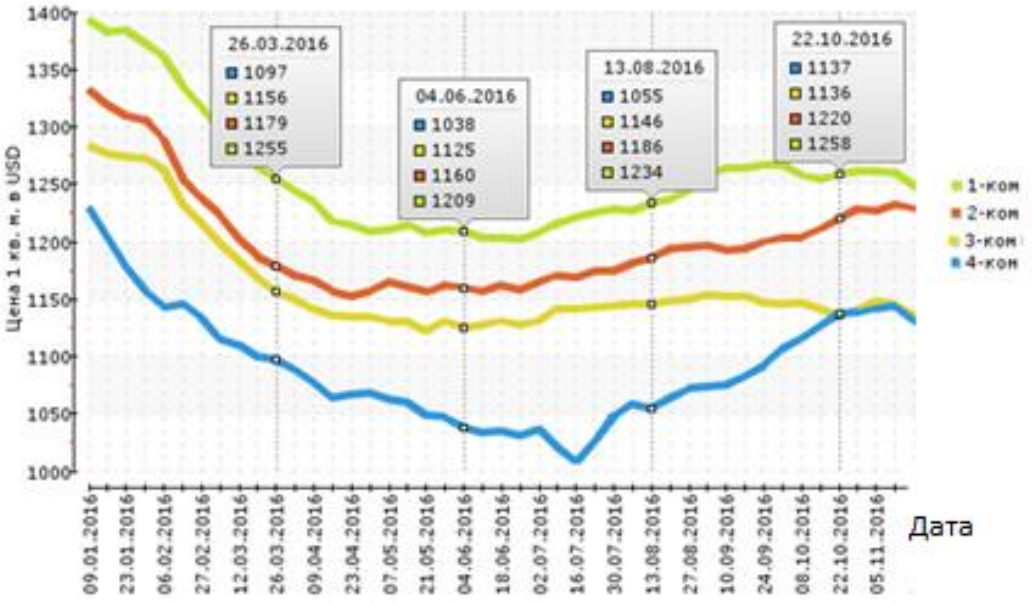


Рис. 2. Цена продажи квартир [3]

Средние цены сделок с квартирами стандартных потребительских качеств в ноябре сформировались на уровне 1.140$ за кв. м. для однокомнатных квартир, 1.050$ за кв. м. для двухкомнатных и 960$ за кв. м. для трехкомнатных квартир.

Так как объем анализируемых данных достаточно велик и выборочные данные цены квартиры не подчиняются нормальному закону распределения, то целесообразно проводить моделирование стоимости квартир по административным районам г. Минска.

В качестве примера рассмотрим построение эконометрической модели стоимости жилья в Партизанском районе. На основе предложений о продаже была построена выборка, содержащая 267 наблюдений (квартир), из них 84 однокомнатных, 115 двухкомнатных, 53 трехкомнатных и 15 четырехкомнатных квартир. В домах кирпичного типа представлено 140 квартир, в домах панельного типа – 127 квартир.

Предварительный графический анализ позволил выявить логарифмическую форму зависимости между ценой квартиры и общей площадью, как наиболее соответствующую данным. Это обусловлено также необходимостью перехода к безразмерным величинам (рисунок 3).

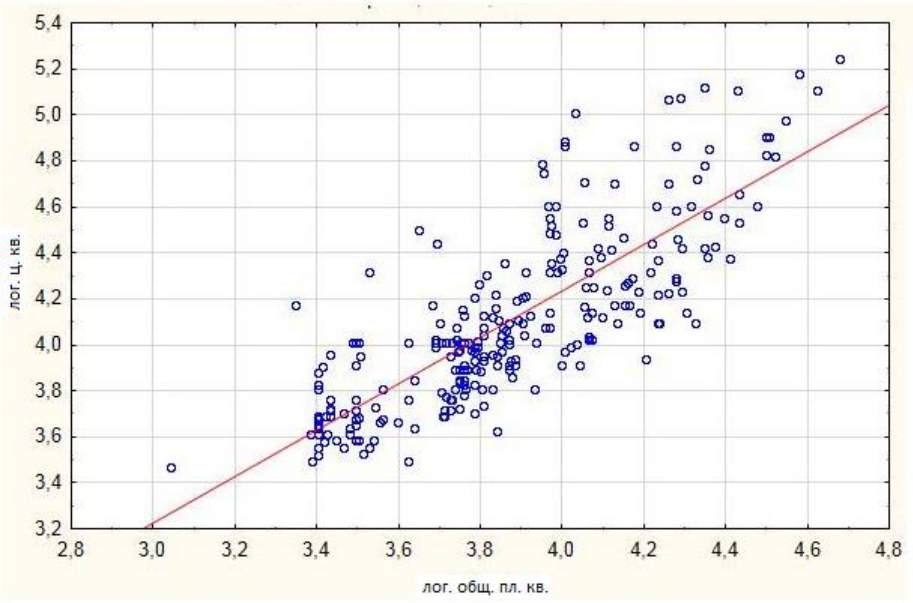


Рис. 3. Логарифмическая зависимость цены от общей площади

Так как в исходных данных нет достоверной информации о состоянии квартиры на данный момент времени, то целесообразно предположить, что чем выше стоимость квадратного метра, тем лучше состояние квартиры [5].

Объекты исследования были разбиты на группы с помощью метода k-средних кластерного анализа на основе значения стоимости квадратного метра квартиры. В результате применения метода наблюдения были разделены на 3 кластера.

В первый кластер вошло 18 квартир. Средняя стоимость квадратного метра равна 2.191$. Минимальная стоимость в кластере 1.990$. Максимальная 2.394$. Квартиры, попавшие в эту группу можно отнести к элитному жилью.

Во второй кластер вошло 86 квартир. Средняя стоимость квадратного метра равна 1.474$. Это квартиры класса «Стандарт». Диапазон цен от 1.346$ до 1.620$

Третий кластер составили квартиры класса «Эконом». Данную группу составляет 161 квартира и средняя стоимость квадратного метра равна 1.100$. Диапазон цен от 987$ до 1.190$.

Для описанных кластеров в модель введены две фиктивные переменные *dv1* и *dv2*.

Эконометрическая модель зависимости стоимости квартиры от типа дома, общей площади и класса квартиры представлена следующим образом:

$$\underset{}{\text{LnPrice}} = \underset{(0,000)}{0,38} + \underset{(0,003)}{0,04}\ \text{Brick} + \underset{(0,016)}{0,97}\ \text{LnTotSp} - \underset{(0,027)}{0,26}\,\text{dv1} + \underset{(0,000)}{0,41}\,\text{dv2} + \varepsilon. \qquad (1)$$

где *LnPrice* - логарифм цены квартиры,

LnTotSp - логарифм общей площади,

Brick - тип дома: 1- дом кирпичный или монолитный, 0 - все остальные.

В модель также включены фиктивные переменные: *dv1*– принимает значение 1, если квартира находится в первом кластере и 0 – в противном случае. Переменная *dv2* - принимает значение 1, если квартира находится в третьем кластере , 0 – в противном случае.

Результаты оценки статистического качества построенной модели (1) представлены в таблице 3.

Все коэффициенты являются статистически значимыми на 5% уровне; множественный коэффициент корреляции $R=0,97$; коэффициент детерминации $R^2=0,95$; отсутствует корре-

ляция в остатках: *DW=1,91, r=0,03*; остатки гомоскедастичны и имеют нормальное распределение: $\chi^2=6,028$, *p=0,1970*. Средний абсолютный процент ошибки *MAPE=0,051*. Данный показатель означает, что модель обладает высокими прогностическими свойствами.

Таблица 3. Критерии качества модели

R	R^2	DW	r_l	F	p_F	χ^2	p_{χ^2}
0.97	0.95	1.91	0.03	1150	0.005	6.028	0.1970

Литература

[1]. Аналитические обзоры компании molnar.by [Электронный ресурс] – Режим доступа: http://molnar.by/analytics/stats/sale_sdel

[2]. Жилищные условия Национальный статистический комитет Республики Беларусь «belstat.gov.by» [Электронный ресурс] – Режим доступа: http://www.belstat.gov.by/ ofitsialnaya-statistika/solialnaya-sfera/zhilischnye-usloviya

[3]. Статистика и аналитика рынка недвижимости Республики Беларусь, информационного каталога realt.by [Электронный ресурс] – Режим доступа: http://realt.by/statistics

[4]. Статистика информационного каталога NB.by [Электронный ресурс] – Режим доступа: http://www.nb.by/statistics/?req=ODYxMWMzNjU3Nz U5ZTJkY mJkNTNmMzUzZjE3YWZlOTY%3D Трифонов, Н.Ю., Шимановский С.А. Эконометрическая модель рынка квартир / Н.Ю. Трифонов, С.А. Шимановский Вопросы оценки. 2002. – № 4. – С. 30 – 35.

МОНИТОРИНГ ФИЗИОЛОГИЧЕСКИХ ПОКАЗАТЕЛЕЙ ЧЕЛОВЕКА ДЛЯ РЕАЛИЗАЦИИ БИОТЕХНИЧЕСКОЙ ОБРАТНОЙ СВЯЗИ В УСТРОЙСТВЕ ИНФРАКРАСНОЙ КАБИНЫ

А.Н. Осипов
Проректор по научной работе БГУИР, кандидат технических наук, доцент

М.М. Меженная
Доцент кафедры инженерной психологии и эргономики БГУИР, кандидат технических наук

М.Х.-М. Тхостов
Старший научный сотрудник Центра 4.13 БГУИР

В.Ю. Драпеза
Магистрантка кафедры электронной техники и технологии БГУИР

Белорусский государственный университет информатики и радиоэлектроники, Республика Беларусь
E-mail: mezhennaya@bsuir.by

Abstract. Energy-efficient mobile infrared (IR) cab has been developed to restore the human body functional state. Infrared cabin provides a deep penetration of optical radiation to human tissue by the use of near-infrared emitters range. Distinctive features of this device are monitored user physiological parameters and automatic control of the parameters of IR procedures based on monitoring results. This allows to generate a thermal stress which adequate to the individual user's functional state.

Введение. Инфракрасное (ИК) излучение используется для проведения тепловых процедур в клинической и спортивной медицине с целью восстановления функциональных резервов человеческого организма. Сеансы ИК терапии сопровождаются рядом позитивных эффектов: расширением кровеносных сосудов, увеличением обмена веществ, усилением иммунитета, повышением содержания кислорода в тканях, тем самым обеспечивая противовоспалительный, противоотечный, противоспазматический и обезболивающий эффекты.

Достигаемый терапевтический эффект воздействия ИК излучения зависит от начального функционального состояния человека и адекватного выбора параметров облучения. Существующие ИК кабины преимущественно воздействуют длинноволновым диапазоном ИК спектра [1-8], способным разогревать только верхние слои кожи без глубокого проникновения в ткани человека [1,9-11]. Кроме того ИК терапия противопоказана при артериальной гипертензии и сердечно-сосудистой недостаточности, так как используемые ИК излучатели генерируют избыточный поток энергии, существенно повышая температуру тела человека. При этом показатели энергопотребления остаются достаточно высокими.

Современный уровень развития технологий позволяет совершенствовать медицинскую технику, в том числе в направлении решения вышеуказанных проблем. При этом перспективной является разработка лечебно-диагностических комплексов с функцией управления параметрами воздействия исходя из физиологических характеристик биообъекта. Применительно к устройствам для инфракрасной терапии это позволит генерировать тепловую нагрузку, адекватную индивидуальному функциональному состоянию пользователя.

В связи с вышеизложенным авторами разработана энергоэффективная мобильная инфракрасная кабина для низкоинтенсивного воздействия ИК излучением преимущественно ближнего ИК диапазона на тело человека. Отличительной особенностью предлагаемого

устройства является реализация биотехнической обратной связи посредством мониторинга физиологических показателей пользователя и автоматического управления параметрами ИК процедуры на основе результатов мониторинга. Задача реализации биотехнической обратной связи может быть успешно решена с применением технологий BigData и нейронных сетей.

Структурная схема и принцип работы устройства. Разработанное авторами устройство для воздействия низкоинтенсивным ИК излучением на человеческий организм представляет собой ИК кабину с автоматическим управлением параметрами воздействия на основе физиологических показателей пользователя (рисунок 1).

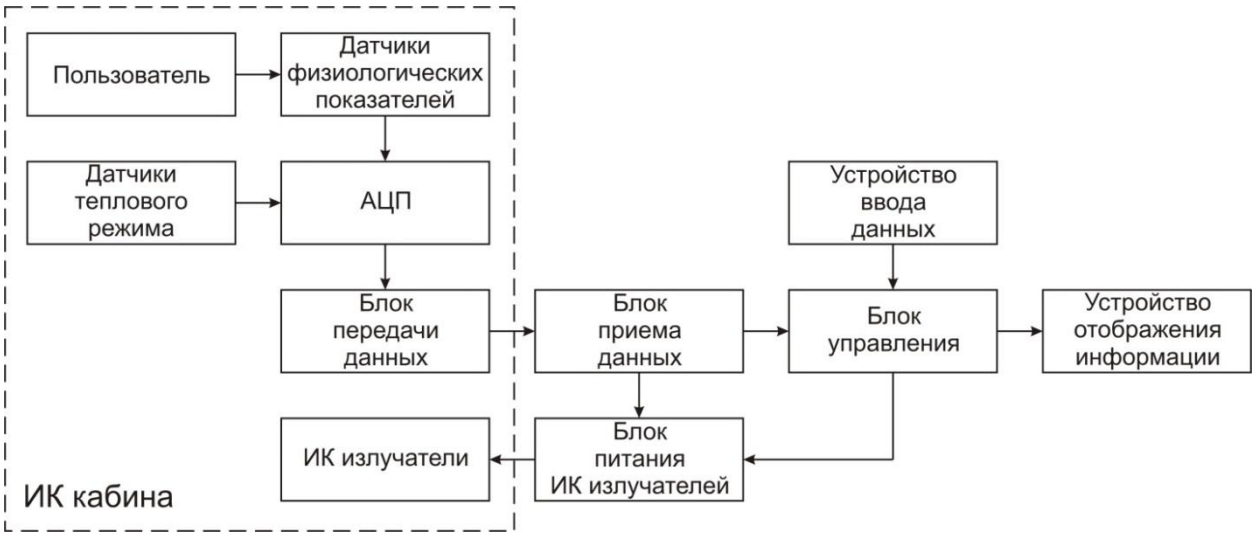


Рис. 1. Структурная схема устройства инфракрасной кабины с автоматическим управлением параметрами воздействия на основе физиологических показателей пользователя

Устройство содержит датчики физиологических показателей пользователя, датчики теплового режима, аналого-цифровой преобразователь (АЦП), блок передачи данных, блок приема данных, блок управления, устройство ввода данных, устройство отображения информации, ИК излучатели, блок питания ИК излучателей.

Устройство функционирует следующим образом.

ИК кабина располагается в вертикальном или горизонтальном положении, включаются ИК излучатели и осуществляется их разогрев до достижения рабочего теплового режима внутри устройства. Контроль теплового режима реализуется посредством датчиков температуры и влажности. Сигналы с датчиков теплового режима преобразуются в цифровую форму посредством АЦП, далее с помощью блоков передачи и приема данных поступают на блок управления, расположенный вне конструкции ИК кабины. С помощью устройства ввода данных устанавливаются требуемые параметры теплового режима. Блок управления осуществляет достижение и поддержание этих рабочих параметров за счет управления блоком питания излучателей. После разогрева ИК излучателей кабина готова к использованию.

Перед началом процедуры на теле пользователя (который предварительно был осмотрен врачом) размещаются датчики физиологических показателей, а именно, сенсоры артериального давления, пульса, температуры тела. Далее пользователь располагается в ИК кабине. Посредством АЦП и блоков передачи и приема данных информация о функциональном состоянии пользователя поступает в блок управления и выводится на устройство отображения в реальном режиме времени, что обеспечивает непрерывное наблюдение за пользователем врачом (оператором).

В процессе проведения терапевтической процедуры осуществляется автоматическая корректировка параметров воздействия на основе мониторинга физиологических показателей

пользователя (биотехническая обратная связь). В частности, посредством управления блоком питания ИК излучателей выполняется регулировка тепловой нагрузки на организм пользователя.

Время процедуры устанавливается посредством блока ввода информации. По истечении требуемого времени терапевтической процедуры происходит автоматическое отключение ИК излучателей блоком управления.

Во время мониторинга физиологических показателей пользователя характер изменения перечисленных биопараметров свидетельствует о происходящих в организме естественных адаптивных процессах терморегуляции. При этом необходимо исключить переход в режим перегрузки и насыщения, критерием наступления которого является превышение вышеуказанными показателями допустимых величин. Для этого целесообразно уменьшать тепловую нагрузку на организм человека посредством снижения мощности ИК излучателей.

Еще одним важным критерием нормального функционирования регуляторных механизмов является появление после начала процедуры быстрой тенденции к восстановлению функциональных показателей. Иная тенденция к восстановлению функциональных показателей является поводом для прекращения ИК процедуры и последующей консультации с врачом.

Дополнительная диагностическая информация о состоянии пользователя может быть получена после окончания процедуры ИК терапии. Это связано с тем, что значения времени для возвращения биопараметров в исходное состояние после окончания ИК процедуры вариабельны у каждого человека (от 5 до 30 минут), но не должны превышать 30 минут. Поэтому предлагаемое устройство реализует возможность контроля физиологических показателей пользователя после окончания процедуры с выводом информации на устройство отображения.

С точки зрения конструктивного исполнения разработанная ИК кабина представляет собой прямоугольную камеру с входной дверью, откидной крышкой для удобства входа в горизонтально расположенную кабину, открывающимися окнами для обеспечения притока воздуха, рефлекторами для защиты головы человека от действия ИК излучения. Внутри ИК кабины размещаются ИК излучатели, датчики тепловой нагрузки, блок АЦП и блок передачи данных. Вне конструкции ИК камеры размещаются блок приема данных, блок управления, устройство ввода данных, устройство отображения информации и блок питания излучателей.

Материал внутренней обшивки кабины – теплоизоляция с зеркальным в ИК диапазоне покрытием из алюминиевой фольги – снижает энергетические затраты и позволяет повысить эффективность прогревания за счет отражения внутренней поверхностью кабины ИК излучения и перенаправления его в центральную зону. Материал внешней обшивки кабины – поликарбонат – предпочтителен с точки зрения дизайна, обеспечивает легкость и мобильность конструкции.

Максимальный физиотерапевтический эффект ИК процедуры достигается за счет использования излучателей ближнего ИК диапазона, которые обеспечивают наибольшую глубину проникновения ИК излучения в ткани человеческого организма [9-11].

Температура воздуха внутри ИК кабины задается посредством изменения мощности электропитания источников ИК излучения и поддерживается на уровне 39°С в области туловища пациента (что существенно ниже существующих серийных аналогов - более 45°С) [2-3] и 32°С в области головы (из-за наличия защитных рефлекторов и воздушных окошек, расположенных на уровне головы). Температура 39°С в области туловища является оптимальной для имитации естественной реакции организма человека на подъем глубокой температуры тела во время развития системного воспалительного процесса и активации при этом защитных нейрогуморальных механизмов. При достижении глубокой температуры тела 39°С у большинства испытуемых обычно не возникает побочных негативных реакций, в первую очередь, со стороны сердечно-сосудистой системы.

Наличие в составе блока питания ИК излучателей понижающего трансформатора обес-

печивает защиту пользователя от случайного поражения электрическим током при проведении терапевтических процедур в случае возникновения неисправности в окружающем оборудовании.

Мониторинг физиологических показателей пользователя при проведении ИК-терапии. Авторами проведены исследования динамики изменения физиологических показателей человека при проведении ИК-терапии посредством вышеописанного устройства.

В исследованиях приняли участие 15 человек (8 мужчин и 7 женщин в возрасте от 19 лет до 31 года). Время сеанса ИК-процедуры составляло 30 минут.

В процессе каждого исследования испытуемый размещался в горизонтально расположенной ИК-кабине. Далее непосредственно в ИК-кабине выполнялась регистрация температуры тела, пульса, верхнего и нижнего артериального давления испытуемого: до начала процедуры, через 15 минут после начала процедуры, через 30 минут после начала процедуры. Для контроля динамики восстановления физиологических показателей после окончания ИК-терапии дополнительно выполнялась регистрация температуры тела, пульса, верхнего и нижнего артериального давления испытуемого: спустя 15 минут, 30 минут и 45 минут после процедуры. Усредненные результаты изменения физиологических показателей с указанием среднеквадратичного отклонения приведены на рисунках 2-5.

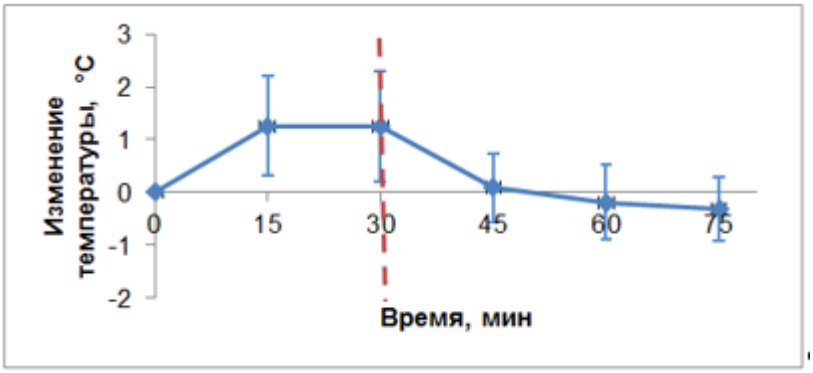


Рис. 2. Среднее арифметическое изменение температуры тела испытуемыхв процессе 30-ти минутного сеанса ИК-терапии, а также в течение 45 минут после окончания ИК-процедуры

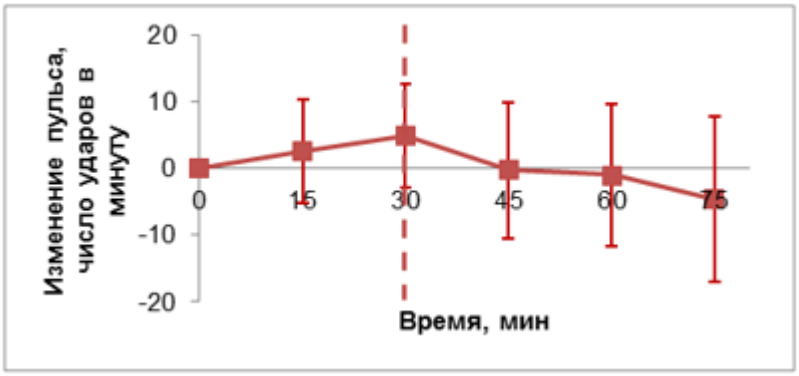


Рис. 3. Среднее арифметическое изменение пульса испытуемых в процессе 30-ти минутного сеанса ИК-терапии, а также в течение 45 минутпосле окончания ИК-процедуры

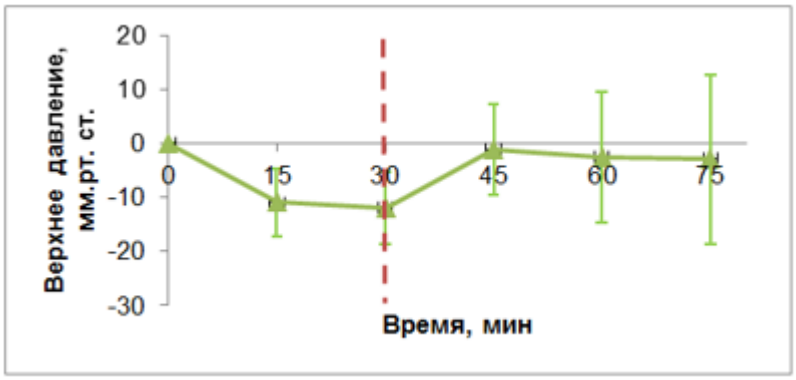


Рис. 4. Среднее арифметическое изменение верхнего артериального давления испытуемых в процессе 30-ти минутного сеанса ИК-терапии, а также в течение 45 минут после окончания ИК-процедуры

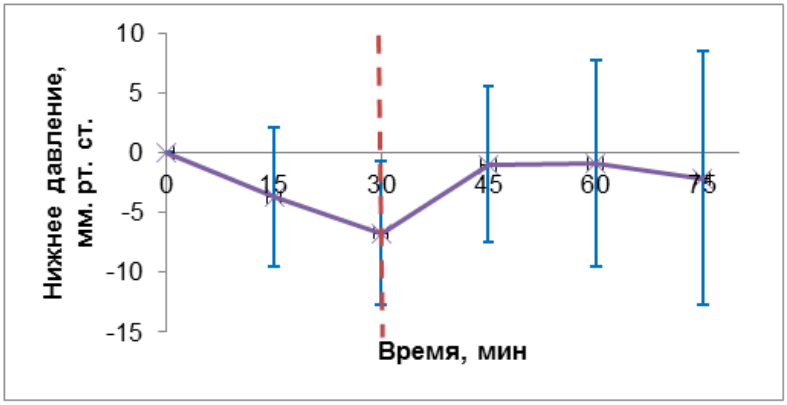


Рис. 5. Среднее арифметическое изменение нижнего артериального давления испытуемых в процессе 30-ти минутного сеанса ИК-терапии, а также в течение 45 минут после окончания ИК-процедуры

Анализ полученных данных выявил следующие закономерности:

1 Температура тела испытуемых в процессе ИК-терапии увеличивалась на 1,24 °С ± 1, 05 °С, спустя 15 минут после окончания процедуры температура возвращалась к первоначальному уровню.

2 Пульс постепенно возрастал в процессе ИК-сеанса и увеличивался в среднем на 4,87 ± 7, 82 удара к моменту окончания процедуры. Далее наблюдалась тенденция к восстановлению исходного уровня уже через 15 минут после завершения процедуры.

3 Особый интерес представляет динамика снижения показателей артериального давления в процессе ИК-терапии. В среднем к моменту окончания процедуры верхнее артериальное давление уменьшалось на 12,00 ± 6,8 мм.рт.ст., ннижнее – на 6,73 ± 6,02 мм.рт.ст. Далее наблюдалась тенденция к восстановлению исходного уровня уже через 15 минут после завершения процедуры.

Полученные результаты позволяют сделать вывод о минимизации тепловой нагрузки на пользователя при проведении ИК терапии посредством разработанного устройства по сравнению с традиционными банями и саунами, а также по сравнению с аналогичными ИК кабинами. Это достигается использованием источников ближнего ИК излучения, а также конструктивными особенностями кабины, позволяющими снизить температуру воздуха при сохранении эффективности прогревания. Наличие защитных рефлекторов и вентиляционных клапанов защищает голову пользователя от нежелательного перегрева. Это в конечном итоге позволяет

расширить сферу применения подобного рода устройств с сугубо бытовой до медицинской за счет устранения ограничений на использование инфракрасных камер при артериальной гипертензии, сердечно-сосудистой недостаточности.

Функция мониторинга физиологических показателей пользователя позволяет получить диагностическую информацию о текущем функциональном состоянии человека. Целью дальнейших исследований авторы видят использование полученной информации для автоматического управления параметрами ИК процедуры, начиная от регулировки температурных режимов и заканчивая полным прекращением процедуры при необходимости. Это позволит адаптировать тепловую нагрузку под индивидуальное функциональное состояние пользователя.

Литература

[1]. Инфракрасные сауны Uborg [Электронный ресурс]. – Режим доступа: http://www.uborgsauna.ru. – Дата доступа : 15.10.2016.

[2]. Воронежский каталог инфракрасных саун и кабин [Электронный ресурс]. – Режим доступа : http://www.iksauna36.ru/manufacturer.php. – Дата доступа : 15.10.2016.

[3]. Сауна и генератор дальнего ИК излучения для нее: пат. WO 2005060355 A2, МПК A61H33/06; опубл. 7.07.2005.

[4]. Дальнее инфракрасное излучение и лучи жизни [Электронный ресурс]. – Режим доступа : http://vitalrays.ru/archives/126. – Дата доступа: 15.10.2016.

[5]. Сауна: пат. 10151789 A1 Германия, МПК A61H33/06; опубл. 30.04.2003.

[6]. Сауна: пат. 3959477 B2 Япония, МПК A61H33/06; опубл. 15.08.2007.

[7]. Infrasun – сауны солнца [Электронный ресурс]. – Режим доступа: http:// http://www.infrasun.ru/infrasun/suncarbon/. – Дата доступа: 15.10.2016.

[8]. Инфракрасные кабины Infradoc [Электронный ресурс]. – Режим доступа: http://www.infradoc.spb.ru/princip.htm. – Дата доступа: 15.10.2016.

[9]. Пономаренко Г.М. Биофизические основы физиотерапии / Г.Н. Пономаренко, И.И. Турковский. М.: "Медицина", 2006. с. 17-18.

[10]. Энциклопедия по охране и безопасности труда / Международная Организация Труда, 2-е изд., 1988.

[11]. Journal of Biomedical Optics 12(4), 044012, 2007.

ЦИФРОВАЯ ОБРАБОТКА РЕЧЕВЫХ СИГНАЛОВ В ДИАГНОСТИКЕ БУЛЬБАРНЫХ НАРУШЕНИЙ

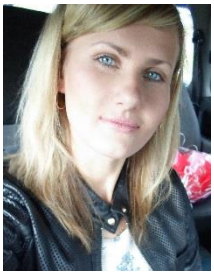
А.Н. Осипов
Проректор по научной работе БГУИР, кандидат технических наук, доцент

С.А. Лихачев[1]
Заведующий неврологическим отделом РНПЦ Неврологии и нейрохирургии, доктор медицинских наук, профессор

Ю.Н. Рушкевич[1]
Ведущий научный сотрудник неврологического отдела РНПЦ Неврологии и нейрохирургии, кандидат медицинских наук, доцен

М.М. Меженная
Доцент кафедры инженерной психологии и эргономики БГУИР, кандидат технических наук

А.А. Борискевич
Профессор кафедры сетей и устройств телекоммуникаций БГУИР,доктор технических наук, доцент



Т.П. Куль
Магистрантка кафедры инженерной психологии и эргономики БГУИР

Белорусский государственный университет информатики и радиоэлектроники, Республика Беларусь
[1] *РНПЦ Неврологии и нейрохирургии МЗ, Республика Беларусь*
E-mail: mezhennaya@bsuir.by, rushkevich@tut.by

Abstract. The method of qualitative and quantitative differential diagnosis of bulbar palsy has been offered on the basis of digital processing of speech signals. The software with the graphic user interface has been developed by authors for implementation of this method which allows to increase the accuracy and speed of diagnosis.

Введение. Бульбарные нарушения представляют собой симптомокомплекс, который включает изменение звучности голоса (дисфонию), затруднения при глотании - дисфагию и замедленность речи, нарушение артикуляции - дизартрию, т. е. симптомы, связанные с вовлечением мускулатуры языка, глотки, гортани и мягкого нёба. Причинами бульбарных нарушений является непосредственное поражение ядер языкоглоточного, блуждающего и подъязычного черепных нервов, расположенных в каудальных отделах ствола головного мозга (бульбарный синдром), а также поражение вышеописанных мышц, нервно-мышечного аппарата, патологические процессы в области ствола мозга и задней черепной ямки [1-3].

К ранним проявлениям бульбарного синдрома относится дисфония: голос больных становится слабым, глухим, истощающимся вплоть до полной афонии. Возникает гнусавость. Звуки при этом произносятся невнятно, «смазанно». Гласные звуки становятся трудноотличи-

мыми друг от друга, согласные звуки, разные по способу образования (твёрдые, мягкие, смычные, щелевые) и месту артикуляции (губные, переднеязычные, заднеязычные), произносятся однотипно с неопределённым местом артикуляции. Речь оказывается резко замедленной и утомляет больных. Развивающаяся дисфагия из-за невозможности сглатывать слюну и приводит к слюнотечению. При бульбарном параличе наступает атрофия мышц языка и выпадают глоточный и нёбный рефлексы. У тяжелобольных с бульбарным синдромом, как правило, развиваются расстройства ритма дыхания и сердечной деятельности, что нередко приводит к смерти [2,3].

Эффективная дифференциальная диагностика позволяет своевременно оказать медицинскую помощь пациентам с бульбарными нарушениями. К достоверным методам диагностики бульбарного синдрома относятся данные электромиографии и прямого осмотра ротоглотки. Однако в настоящее время имеются сложности постановки диагноза на ранней стадии, сопровождающейся, как отмечалось выше, нарушениями речевой функции. Для проведения своевременной и объективной диагностики бульбарных нарушений предлагается использовать методы цифровой обработки речевых сигналов.

Методика регистрации и обработки речевых сигналов. Диагностические исследования бульбарных нарушений были проведены на базе РНПЦ неврологии и нейрохирургии МЗ РБ. В группе пациентов с боковым амиотрофическим склерозом с бульбарным синдромом, а также в контрольной группе здоровых лиц были зарегистрированы тестовые речевые сигналы. Тест представлял собой счет от одного до десяти.

Последующая обработка речевых сигналов выполнялась в среде MatLab с помощью специально разработанного авторами статьи программного обеспечения с графическим интерфейсом. Обработка включала следующие этапы:

1 Автоматическое выделение в зарегистрированном сигнале речевых фрагментов.
2 Подсчет количества выделенных речевых фрагментов.
3 Построение спектрограммы зарегистрированного сигнала.
4 Построение кепстрограмм для выделенных речевых фрагментов.
5 Определение с помощью кепстральной функции частоты основного тона для каждого речевого фрагмента.
6 Расчет средней величины частоты основного тона.
7 Расчет коэффициента вариации частоты основного тона.
8 Построение гистограммы для массива, представляющего собой результат «склейки» всех выделенных речевых фрагментов.
9 Вычисление средней амплитуды выделенных речевых фрагментов.
10 Вычисление общего времени всех выделенных речевых фрагментов.
11 Вычисление коэффициента асимметрии гистограммы.
12 Вычисление коэффициента эксцесса гистограммы.

Далее приведена подробная методика реализации вышеописанных этапов.

Исходный зарегистрированный сигнал характеризовался частотой дискретизации 44,1 кГц, разрядностью 16 бит. Предварительно производилось усреднение зарегистрированного сигнала в окне без перекрытия для снижения исходной частоты дискретизации:

$$A[i] = \frac{1}{M}\sum_{j=0}^{M-1} x[i \cdot M + j] \qquad (1)$$

где $A[i]$ – отсчеты, полученные из исходного сигнала x посредством усреднения;

$i = 0...\frac{N}{M} - 1$ – номер окна; N – число отсчетов исходного речевого сигнала x;

M – длина окна (число усредняемых точек);

j – номер временного отсчета внутри окна.

В результате усреднения при $M = 5$ частота дискретизации была понижена до 8,82 кГц. Это позволило впоследствии увеличить скорость обработки данных без потери полезной информации в сигнале.

Речь человека содержит паузы между словами. Традиционно для решения задачи разделения речевого сигнала на голосовые и неголосовые участки исходный сигнал разделяется на фрагменты длиной 5-100 мс (с точки зрения динамики речи самые быстрые изменения могут происходить всего за несколько миллисекунд, в то время, как некоторые гласные звуки остаются относительно стабильными в течение 100-200 мс). Для классификации принадлежности фрагмента к сигналу или паузе рассчитывалась кратковременная энергия сигнала в данном фрагменте:

$$E_m = \sum_{j=0}^{L_{fr}-1} A\left[m \cdot L_{fr} + j\right]^2 \tag{2}$$

где L_{fr} – длина фрагмента;

$m = 0\,..\frac{N}{M \cdot L_{fr}} - 1$ – количество фрагментов; j – номер временного отсчета усредненного сигнала внутри фрагмента. В качестве L_{fr} авторами выбраны 400 отсчетов, что соответствует временной реализации сигнала в 45,4 мс.

На основе экспериментальных исследований речевых сигналов в норме было сформировано условие, при выполнении которого принималось решение о принадлежности *m*-ого фрагмента к речи:

$$E_m \geq level_E \cdot \langle E \rangle, \tag{3}$$

где $\langle E \rangle$ – средняя кратковременная энергия всех фрагментов, $level_E$ – пороговый уровень кратковременной энергии.

Авторами установлено, что при $level_E = 0.2$ происходит автоматическое выделение слов и/или отдельных фонем в сигнале.

Далее выполнялось построение спектрограммы сигнала. Для этого речевой сигнал разделялся на временные отрезки, в пределах которых его можно считать стационарным (5-100 мс). Исходный сигнал A на выбранном отрезке умножался на оконную функцию w и подвергался быстрому преобразованию Фурье в соответствии с выражением:

$$STFT_A^w\left[f_k, \tau\right] = \sum_{i=0}^{L-1} A[i] \cdot w[i] \cdot e^{-\frac{j2\pi ki}{L}},\ f_k = \frac{k \cdot f_d}{L}, k = 0..(L-1)/2, \tag{4}$$

где L – длина окна, τ – величина перекрытия окон, f_d – частота дискретизации.

После данной операции путем возведения в квадрат амплитудной части оконного преобразования Фурье получали спектрограмму мощности для анализируемого окна:

$$Спектрограмма\ A\left[f, t\right] = \left|STFT_A^w\left[f_k, \tau\right]\right|^2. \tag{5}$$

Далее производилось смещение окна на величину τ и процедура повторялась. Подобным образом анализировались все подинтервалы сигнала и строилась результирующая

спектрограмма, представляющая собой двумерную матрицу, строки которой соответствуют временным отсчетам t от 0 секунд до окончания времени регистрации речевого сигнала, столбцы – частотам f от 0 до 4,41 кГц, а в ячейках рассчитана амплитуда сигнала [4]. В качестве основных параметров частотно-временной обработки выбраны следующие: окно Хэмминга, размер окна L в 512 отсчетов, частота дискретизации f_d в 8,82 кГц, перекрытие окон τ в 50%. Указанные характеристики обеспечивают качественное частотно-временное представление речевого сигнала, высокое разрешение по частоте Δf = 17,2 Гц и по времени Δt = 29,0 мс:

$$\Delta f = \frac{L \cdot (100 - \tau)}{f_d \cdot 100}, \quad (6)$$

$$\Delta t = \frac{f_d}{L}. \quad (7)$$

Для определения частоты основного тона сигнала использовался метод определения кепстра, заключающийся в применении к модулю спектральной плотности исследуемого сигнала обратного преобразования Фурье. При этом в кепстрограмме вокализованного отрезка звука появляется пик на расстоянии основного тона сигнала, что и является основополагающим для последующего вычисления частоты основного тона.

Частота основного тона вычислялась для каждого выделенного вокализованного фрагмента сигнала. По итогам расчетов определяли среднее значение частоты основного тона, а также коэффициент вариации данного параметра – относительную меру разброса значений признака в статистической совокупности. Значения коэффициента вариации менее 10 % свидетельствуют о малом рассеянии, от 10 % до 20 % – о среднем рассеянии, более 20 % – о сильном рассеянии вариант относительно средней арифметической величины.

Для расчета статистических показателей выполнялась «склейка» всех выделенных речевых фрагментов в единый массив. Для полученного массива рассчитывалась средняя амплитуда, а также длительность, соответствующая общей продолжительности речи.

Для визуализации данных на этапе статистической обработки выполнялось построение гистограммы для массива всех речевых фрагментов. Далее для оценки однородности распределения данных в речевых фрагментах рассчитывались показатели асимметрии и эксцесса гистограммы.

Коэффициент асимметрии может быть положительным (для правосторонней асимметрии) и отрицательным (для левосторонней асимметрии). Асимметрия выше 0,5 (независимо от знака) считается значительной, меньше 0,25 – незначительной.

Показатель эксцесса отражает, насколько резкий скачок имеет изучаемое явление. Если показатель эксцесса больше нуля, то распределение островершинное и скачок считается значительным, если коэффициент эксцесса меньше нуля, то распределение считается плосковершинным и скачок считается незначительным.

Результаты исследований. Результаты обработки тестовых речевых сигналов в норме, при бульбарном синдроме до лечения и после курса транскраниальной магнитной стимуляции и нейрометаболического лечения представлены на рисунках 1, 2 3 соответственно.

Анализ полученных результатов выявил следующие закономерности.

В группе здоровых лиц (рис. 1) количество распознанных речевых фрагментов составляет 10-12, что соответствует количеству произносимых слов (10) или фонем («че-тыре», «во-семь). На спектрограмме отчётливо выделяются равностоящие друг от друга речевые фрагменты, частота основного тона, а также формантные частоты. Кепстрограммы для распознанных речевых фрагментов также имеют характерные пики в области основного тона и кратных ему формантных частот. Коэффициент вариации основного тона невысокий (равен 7 на рис.1), что свидетельствует о постоянстве данного параметра во время речи. Гистограмма речи характеризуется симметричностью (коэффициент асимметрии равен 0,58 на

рис.1) с четко выделенным пиковым значением (коэффициент эксцесса равен 12,39 на рис.1).

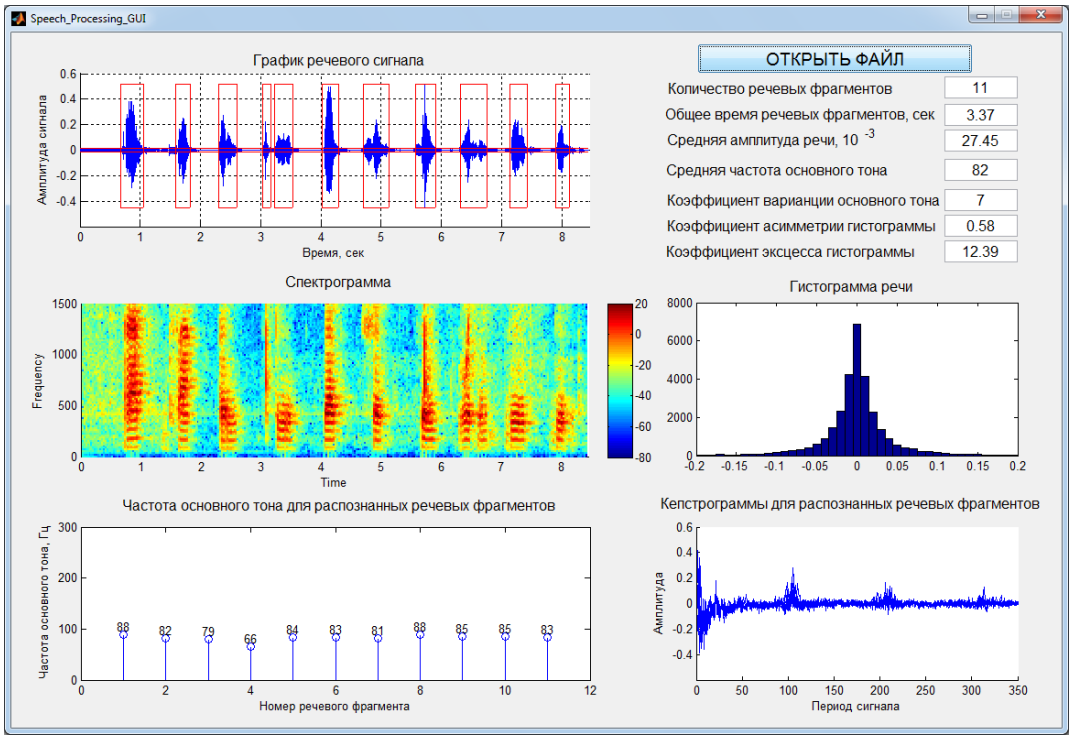


Рис. 1. Результаты обработки речевого сигнала в норме (испытуемый У)

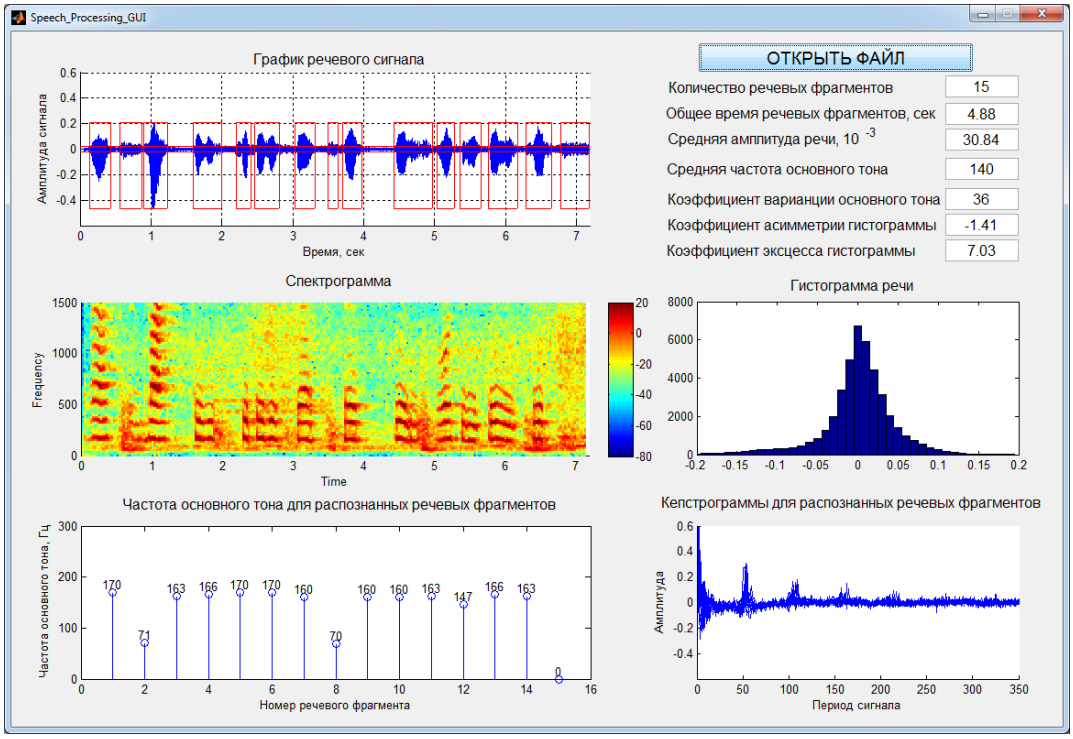


Рис. 2. Результаты обработки речевого сигнала
до лечения бульбарного синдрома (пациент К.)

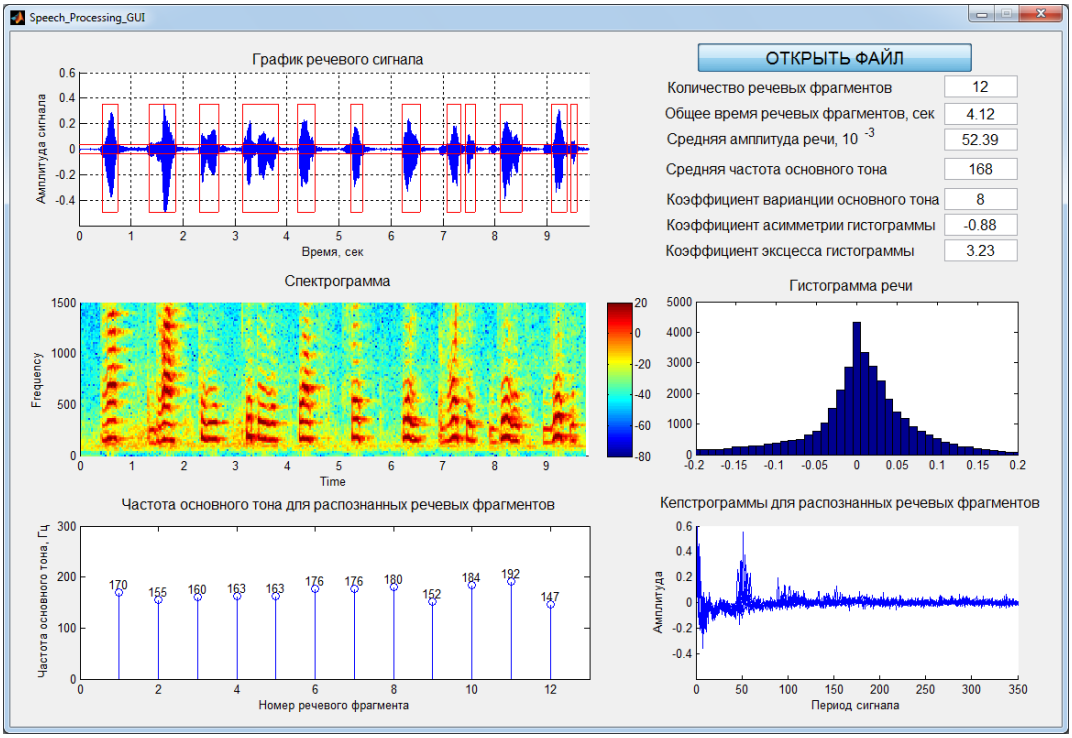


Рис. 3. Результаты обработки речевого сигнала после лечения бульбарного синдрома (пациент К.)

В группе пациентов с бульбарным синдромом до лечения (рис. 2) количество распознанных речевых фрагментов, как правило, превышает количество произносимых слов (10). Это объясняется характерной для данной патологии невнятностью (гнусавостью) речи. Сигнал на спектрограмме не имеет четкой временной структуры в виде равноотстоящих речевых актов, как у здоровых лиц. Значение частоты основного тона для ряда речевых фрагментов не входит в стандартные диапазоны 70 – 450 Гц или не определяется вовсе. Показателен коэффициент вариации частоты основного тона (равен 36 на рис.2), свидетельствующий о сильной степени рассеяния данного параметра относительно среднеарифметического значения. Гистограмма характеризуется левосторонней асимметрией (коэффициент асимметрии равен -1,41 на рис.2), менее выраженным пиковым значением по сравнению с нормой (коэффициент эксцесса равен 7,03 на рис.2).

В группе пациентов с бульбарным синдромом после лечения (рис. 3) количество распознанных речевых фрагментов в целом соответствовало количеству произносимых слов (10) или фонем («че-тыре», «во-семь). Возросла амплитуда сигнала (см. рис.2,3). На спектрограмме речевые фрагменты приобрели четкие очертания; выделяются паузы, как в группе здоровых лиц. Характерные для основного тона и формантных частот пики демонстрируют кепстрограммы. Вариабельность частоты основного тона вернулась к показателям в норме (равна 8 на рис.3). Степень асимметричности гистограммы снижается (коэффициент асимметрии равен -0,88 на рис.3).

Заключение. Предложен метод качественной и количественной диагностики бульбарных нарушений на основе цифровой обработки речевых сигналов. Для реализации данного метода авторами разработано программное обеспечение с графическим интерфейсом, которое позволяет повысить точность и скорость постановки диагноза.

Установлено, что речевые сигналы пациентов с бульбарным синдромом содержат число вокализованных фрагментов, превышающее количество произносимых слов (из-за невнятности речи); сигнал на спектрограмме не имеет четкой временной структуры в виде равноотстоящих речевых актов, как у здоровых лиц; значение частоты основного тона для ряда речевых

фрагментов не входит в стандартные диапазоны 70 – 450 Гц или не определяется вовсе; наблюдается высокая степень рассеяния значений частоты основного тона; гистограмма речи асимметрична.

Метод цифровой обработки речевых сигналов также целесообразно использовать для контроля эффективности лечения неврологических патологий, сопровождающихся нарушениями речевой функции. Установлено, что речевые сигналы пациентов с бульбарным синдромом после лечения характеризуются увеличенной амплитудой, снижением вариабельности частоты основного тона, а также степени асимметричности гистограммы речи. На спектрограмме речевые фрагменты приобретают четкие очертания; выделяются паузы, характерные для основного тона и формантных частот пики демонстрируют кепстрограммы.

Литература

[1]. Завалишин, И.А. Боковой амиотрофический склероз / И.А. Завалишин – М.: ГЭОТАР-Медиа, 2009: 272.

[2]. Andersen, P. EFNS guidelines on the clinical management of amyotrophic lateral sclerosis (MALS)-revised report of an EFNS task force / P. Andersen, et al. Eur J Neurol. 2012;19(3):360–75.

[3]. Miller, R. Practice parameter update: the care of the patient with amyotrophic lateral sclerosis: drug, nutritional, and respiratory therapies (an evidence-based review): report of the Quality Standards Subcommittee of the American Academy of Neurology / R. Miller, et al. Neurology. 2009;73(15):1218–1226. doi: 10.1212/WNL.0b013e3181bc0141.

[4]. Райгайян, Р.М. Анализ биомедицинских сигналов. Практический подход / Р.М. Райгайян – М.: ФИЗМАТЛИТ, 2007. – 440 с.

ПРОБЛЕМА РАСПОЗНАВАНИЯ ИМЕНОВАННЫХ СУЩНОСТЕЙ В БИОМЕДИЦИНСКИХ ПУБЛИКАЦИЯХ

А. В. Пашук
Ассистент кафедры информатики, магистр технических наук

А. Б. Гуринович
Доцент кафедры вычислительных методов и программирования, кандидат физико-математических наук, доцент

Н.А. Волорова
Заведующая кафедой информатики БГУИР, кандидат технических наук, доцент

А. П. Кузнецов
Проректор по научной работе, доктор технических наук, профессор

Белорусский государственный университет информатики и радиоэлектроники, Республика Беларусь
E-mail: pashuk@bsuir.by

Abstract. The number of publications in biomedicine published and indexed annually by PubMed [1] almost doubled over the past 10 years (from 746 thousand to 1354 thousand). This leads to a deterioration in the quality of search and cataloging of scientific publications and it becomes increasingly difficult for scientists to find the necessary information. There is a need to transform unstructured scientific texts into structured formats (XML, JSON). In this task, the quality of recognition of named entities in textual information is of great importance.

Распознавание именованных сущностей (Named-Entity Recognition, сокращенно NER) – это одна из задач извлечения информации из неструктурированного текста, которая заключается в поиске и классификации именованных сущностей, таких как люди, организации, протеины, гены и т.д. Существует целый ряд различных решений данной задачи, большинство из которых позволяет добиться 90% величины F-меры (метрика оценки качества классификатора, объединяющая другие метрики – полноту и точность). Однако, сказанное выше применимо только к классификации ненаучных источников информации. Научная литература и, в частности, биомедицинские статьи и публикации имеют ряд особенностей, которые не позволяют получить хороших результатов с использованием классические алгоритмы распознавания:

– большое количество сокращений, которые могут принимать различные значения, в зависимости от контекста;

– различные варианты написания терминов (например, TNFα, TNF α или TNF-α);

– большое количество синонимов (например, NaCl (соль) имеет больше 300 синонимов [5]);

– наличие двоякости (так называемые омографы), например, CAT (ген) и cat (кошка);

– корректное определение границ термина.

Результат работы классификатора можно представить матрицы неточностей, приведенных в таблице 1.

Зная значения данной матрицы можно посчитать полноту (recall) и точность (precision) поиска – основные метрики, позволяющие оценить качество работы алгоритма поиска.

Таблица 1. Матрица ошибок

		Размеченное значение	
		Правильно	Неправильно
Реальное значение	Правильно	True Positive (TP)	False Negative (FN)
	Неправильно	False Positive (FP)	True Negative (TN)

$$recall = \frac{TP}{TP + FP} \tag{1}$$

$$precision = \frac{TP}{TP + FN} \tag{2}$$

Чтобы улучшить качество работы классификатора именованных сущностей используются различные правила нормализации. Для оценки качества нормализации используются две основных метрики: двоякость (ambiguity) и вариативность (variability) терминов.

Если представить словарь как список терминов $\{t_1, \dots, t_N\}$, где каждый термин связан с идентификатором понятия $c_j \in \{c_1, \dots, c_M\}$. В этом случае двоякость и вариативность терминов можно выразить следующими выражениями [1]:

$$ambiguity = \frac{1}{N}\sum_{i=1}^{N} C(t_i), \tag{3}$$

где N - количество терминов в словаре (онтологии);

$C(t_i)$ - количество понятий в словаре, которые включают в себя слова, по написанию совпадающие с термином t_i.

$$variability = \frac{1}{M}\sum_{j=1}^{M} T(t_j), \tag{4}$$

где M - количество понятий в словаре (онтологии);

$T(T_j)$ - количество уникальных терминов, которые включает понятие c_j.

В таблице 2 приведен пример термина, определенного в нескольких онтологиях (CheBI, DrugBank и др.) и имеющего различные значений в каждой из онтологий.

Нормализацию необходимо использовать только в случае, если она уменьшает одну из метрик (1) или (2), при этом не увеличивая другую. Также показателем качества нормализации служат увеличение точности (precision) и полноты (recall) поиска. Улучшить эти метрики можно, уменьшив количество ошибок первого рода (false positives) или ошибок второго рода (false negatives).

Одним из методов нормализации, позволяющим уменьшить количество ошибок второго рода является использование так называемых генераторов вариантов терминов (Term Variant Generator, сокращенно TVG). Такие генераторы позволяют учитывать не только документы из онтологии, но также различные варианты эти терминов. Например, аббревиатура протеина TNFα может быть записана как TNF α или TNF-α, при этом в онтологии как правило существует один вариант написания (в некоторых случаях онтологии имеют списки синонимов или популярные варианты написания понятия). С использованием TVG будут сгенерированы дополнительные варианты для поиска в онтологии, согласно заложенным правилам.

Таблица 2. Пример термина, имеющего несколько значений

Термин	Онтология	Совпадение в онтологии (ID)	Синонимы термина из онтологии
IMP	ChEBI	Inosine Monophosphate (D007291)	ribosylhypoxanthine monophosphate, inosinic acid, **IMP**, inosinate, sodium, sodium inosinate, inosine monophosphate, acids, inosinic, monophosphate, ribosylhypoxanthine, inosinic acids, monophosphate, inosine, acid, inosinic
	DrugBank	Imipenem (DB01598)	imipemide, imipenem anhydrou, imipenem anhydrous, n-formimidoylthienamycin, imipenem, **IMP**, imipenem and cilastatin for injection, USP, ran-imipenem-cilastatin, imipenem, n-formimidoyl thienamycin, imipenem and cilastatin, imipenemum, imipenem and cilastatin for injection, -USP, primaxin 250, imipenem and cilastatin for injection USP, (5R,6S)-6-((R)-1-Hydroxyethyl)-3-(2-(iminomethylamino) ethylthio)-7-oxo-1-azabicyclo(3.2.0) hept-2-ene-2-carbonsaeure, primaxin IV 500, primaxin 500, primaxin IV 250/250 add-vantage vial, imipenem and cilastatin for injection, usp, imipenem and cilastatin for injection,-usp, (5R,6S)-3-(2-formimidoylamino-ethylsulfanyl)-6-((R)-1-hydroxy-ethyl)-7-oxo-1-aza-bicyclo[3.2.0] hept-2-ene-2-carboxylic acid, imipenem and cilastatin for injection, tienamycin, imipenem and cilastatin for injection usp, imipenem and cilastatin for injection-USP, imipenem and cilastatin for injection-usp, primaxin IV, primaxin-iv, n-formimidoyl thienamycin, (5R,6S)-3-((2-(formimidoylamino) ethyl) thio)-6-((R)-1-hydroxyethyl)-7-oxo-1-azabicyclo(3.2.0) hept-2-ene-2-carboxylic acid
	GeneOntology	obsolete mitochondrial inner membrane peptidase activity (GO:0004244)	**IMP**, obsolete mitochondrial inner membrane peptidase activity, mitochondrial inner membrane peptidase activity
		mitochondrial inner membrane peptidase complex (GO:0042720)	**IMP**, mitochondrial inner membrane peptidase complex
	ChEBI	IMP (CHEBI:17202)	**IMP**, C10H13N4O8P
	Uniprot (Для нескольких организмов)	Inositol monophosphatase (IMPA1_DICDI, IMPP_MESCR)	IMPase, **IMP**, inositol-1(or 4)-monophosphatase, inositol monophosphatase, d-galactose 1-phosphate phosphatase

Рассмотренный в [4] генератор вариантов предусматривает создание вариантов по следующим правилам: преобразование термина во множественный/единичный вид, удаление/добавление знаков пунктуации (таких как дефисы или апострофы). В рамках исследования был разработан собственный алгоритм, расширяющий список правил стандартного генератора для лучшего определения биомедицинских терминов:

– преобразование чисел в начале термина (5-iodotubercidin, 5iodotubercidin, 5 iodotubercidin и т.д.);

– преобразование числе в конце термина (IL-1, IL 1, IL1 и т.д.);

– буквенные индексы в конце термина (penicillin G, penicillin-G и т.д.);

– преобразование греческих символов в их текстовой выражение (TNF α, TNF alpha) и др.

Гистограммы количества оригинальных вариантов (синонимы, варианты написания и т.д.) и количества сгенерированных вариантов (включая оригинальные) изображены на рисунке 1.

Описанные выше правила позволяют классификатору корректно определить и разметить термины в биомедицинской статье или публикации. Также стоит отметить, что из всей массы полученных вариантов в публикациях обычно встречается не больше 30%. Поэтому после генерации вариантов полученный алгоритм проверяет их на текстах статей, отбрасывая варианты, который ни разу не встречаются. Такой подход позволяет уменьшить время и нагрузку на систему, затрачиваемые на разметку текста.

Отдельно стоит рассмотреть вопрос влияния генераторов вариантов на двоякость и вариативность. В рамках моделирования были рассмотрены онтологии, содержащие классификации генов (GeneOntology [2]) и болезней (ICD-10 [3]). Были посчитаны двоякость (рисунок 1) и вариативность (рисунок 2) терминов в данных онтологий с и без использования генераторов вариантов.

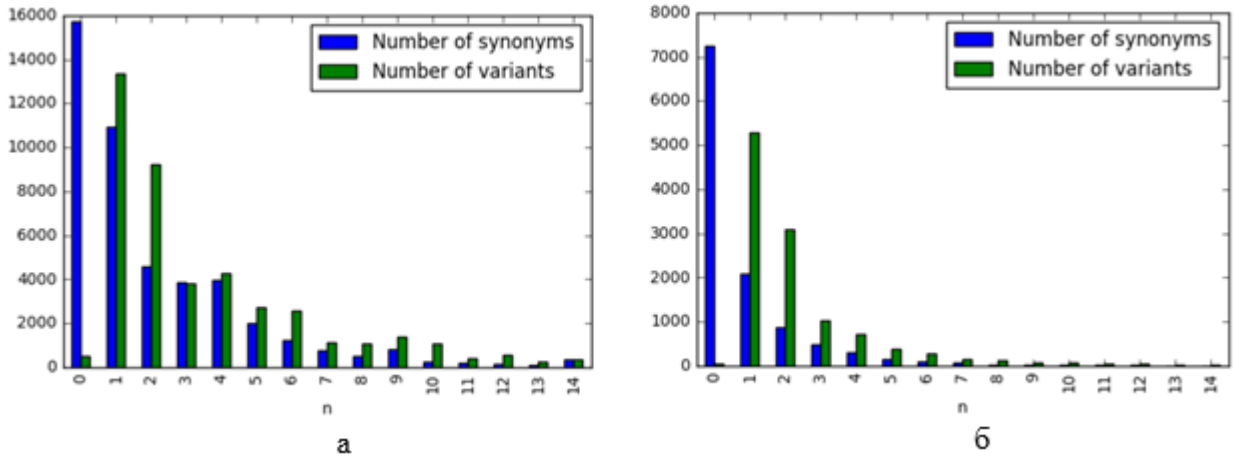


Рис. 1. Количества вариантов написания терминов в онтологиях GeneOntology (а) и ICD-10 (б) с и без использования TVG

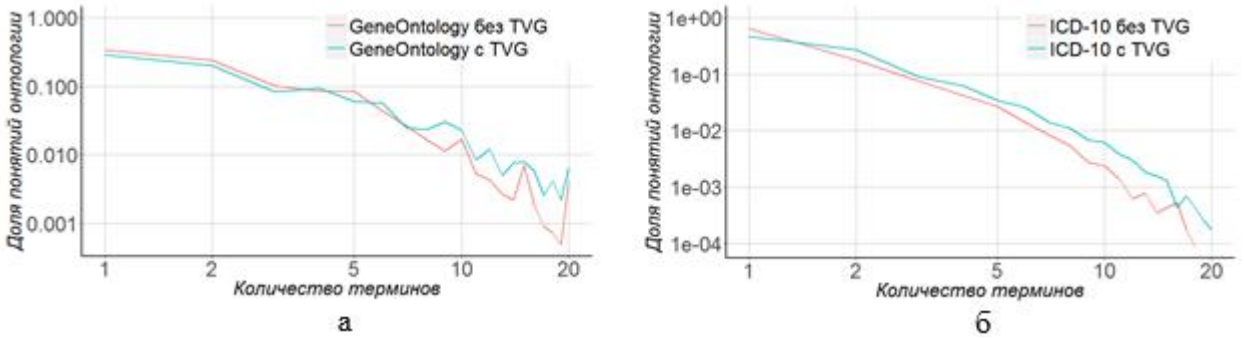


Рис. 2. Изменение двоякости терминов в онтологиях GeneOntology (а) и ICD-10 (б) с и без использования TVG

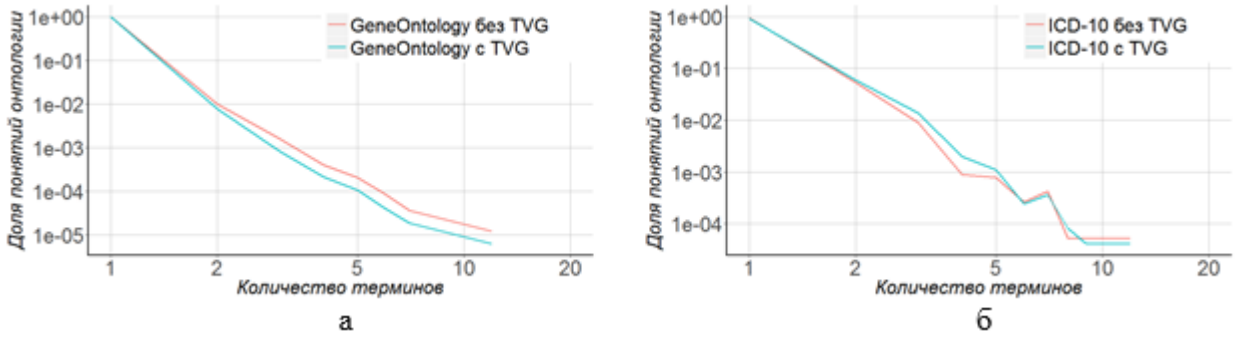


Рис. 3. Изменение вариативности терминов в онтологиях GeneOntology (а) и ICD-10 (б) с и без использования TVG

Из рисунков 2-3 видно, что внедрение генератора не оказало существенного влияния на данные метрики. В то же время, внедрение позволило увеличить полноту поиска (в среднем около 24%), при этом практически не уменьшив его точность (1-2%). Эксперименты показывают, что использование дополнительных вариантов терминов значительно увеличивает шум (количество корректных терминов, для которых найдены некорректные совпадения в онтологиях при разметке). Данную проблему можно решить с помощью внедрения дополнительных фильтров, учитывающих контекст, в котором встречается термин, что является следующим этапом исследования.

Литература

[1]. 1. PubMed NCBI // National Center for Biotechnology Information [Electronic resource]. - 2017. - Mode of access: https://www.ncbi.nlm.nih.gov/pubmed. - Date of access: 13.03.2017.

[2]. 2. Tsuruoka, Y. Normalizing biomedical terms by minimizing ambiguity and variability / Y. Tsuruoka, J. McNaught, S. Ananiadou // BMC Bioinformatics 2008, 9 (Suppl 3) [Electronic resource]. - Mode of access: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-9-S3-S2. - Date of access: 15.03.2017.

[3]. 2. International Statistical Classification of Diseases and Related Health Problems 10th Revision // World Health Organization [Electronic resource]. - 2016. - Mode of access: http://apps.who.int/classifications/icd10/browse/2010/en. - Date of access: 15.03.2017.

[4]. 3. Gene Ontology Consortium // Gene Ontology Consortium (GOC) [Electronic resource]. - 2015. - Mode of access: http://www.geneontology.org. - Date of access: 14.03.2017.

[5]. 4. Tsuruoka, Y. Probabilistic term variant generator for biomedical terms / Y. Tsuruoka, J. Tsujii: SIGIR '03 Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval [Electronic resource]. - Mode of access: http://www.nactem.ac.uk/tsuruoka/papers/sigir03.pdf. - Date of access: 15.03.2017.

[6]. 5. Sodium Chloride // Pubchem: Open Chemistry Database [Electronic resource]. - 2017. - Mode of access: https://pubchem.ncbi.nlm.nih.gov/compound/sodium_chloride#section=MeSH-Synonyms. - Date of access: 13.03.2017.

ИССЛЕДОВАНИЕ ВОЗДЕЙСТВИЯ ФИЗИОТЕРАПЕВТИЧЕСКИХ ФАКТОРОВ НА МИКРОЦИРКУЛЯЦИЮ ПОВЕРХНОСТНЫХ БИОТКАНЕЙ ЧЕЛОВЕКА

С.К. Дик
Первый проректор БГУИР, кандидат физико-математических наук, доцент

Т.В. Гордейчук
Ассистент кафедры инженерной психологии и эргономики, аспирант БГУИР, магистр технических наук

М.М. Меженная
Доцент кафедры инженерной психологии и эргономики БГУИР, кандидат технических наук

С.Н. Табунов[1]
Главнй врач Санатория«Лесное»

Г.Д. Ситник[2]
Заместитель директора по организационно-методической работе РНПЦ неврологии и нейрохирургии, кандтдат медицинских наук, доцент

П.И. Никитенко
Студентка кафедры электронной техники и технологии БГУИР

Е.Н. Рункевич
Студентка кафедры электронной техники и технологии БГУИР

И.В. Кишкевич
Студентка кафедры электронной техники и технологии БГУИР

Белорусский государственный университет информатики и радиоэлектроники, Республика Беларусь
[1]Санаторий «Лесное», Республика Беларусь
[2]Республиканский научно-практический центр неврологии и нейрохирургии, Республика Беларусь
E-mail: t.gordeychuk@bsuir.by

Abstract. The number of publications in biomedicine published and indexed annually by PubMed [1] almost doubled over the past 10 years (from 746 thousand to 1354 thousand). This leads to a deterioration in the quality of search and cataloging of scientific publications and it becomes increasingly difficult for scientists to find the necessary information. There is a need to transform unstructured scientific texts into structured formats (XML, JSON). In this task, the quality of recognition of named entities in textual information is of great importance.

Введение. Мониторинг состояния системы микроциркуляции как основного звена, обеспечивающего метаболический гомеостаз в органах и тканях, является одной из важных проблем современной медицинской диагностики, так как функциональные и морфологические изменения в микроциркуляторном русле наблюдаются при многих заболеваниях: сердечно-

сосудистых осложнениях, атеросклерозе, сахарном диабете, хронической венозной недостаточности и других [1]. В настоящее время мониторирование микроциркуляционной функции ограниченно по ряду причин, основными из которых являются: существование ограниченного числа безопасных методов исследования и сложность интерпретации получаемых данных.

Для изучения системы микроциркуляции все чаще применяются оптические методы диагностики, обладающие следующими преимуществами: высокой точностью и чувствительностью, дистанционностью, высоким пространственным разрешением и воспроизводимостью результатов измерений [2]. По сравнению с традиционно используемыми в медицинской практике морфологическими исследованиями, проводящимися в большинстве случаев биопсийным методом, отражающими состояние микроциркуляции только в конкретной точке и не дающие представлений о динамических процессах, данные методы характеризуются неинвазивностью и безопасностью для пациента [2,3]. Возможность проведения диагностики состояния сосудистой системы и микроциркуляции крови в режиме реального времени обеспечивается рядом оптических методов: лазерная допплеровская флуометрия, доплеровская оптическая когерентная томография, интравитальная микроскопия, магнитнорезонансная томография и ангиография, транскраниальная доплерография, лазерная спекл-визуализация и др. Однако, некоторые из них имеют ряд существенных ограничений: недостаточно высокое пространственное и временное разрешение, ограниченность информации о потоке частиц, особенно при сканировании по глубине биоткани, инвазивность измерений и др. [4]

Основная часть. Оптические методы являются перспективным инструментом диагностики и лечения заболеваний человека вследствие присущих им преимуществ: бесконтактность, высокая точность и чувствительность, дистанционность, высокое пространственное разрешение и воспроизводимость результатов измерений [1,2].

Использование оптических методов для исследования кожного покрова человека позволяет оценить состояние биологических тканей на различной глубине и с различной разрешающей способностью. По сравнению с традиционно используемой в медицинской практике биопсией данные методы характеризуются неинвазивностью и безопасностью для пациента. При этом большинство современных оптических методов (дерматоскопия, оптический видеомонитроинг, оптическая топометрия, 3D-моделирование кожи, оптическая когерентная томография) нацелены на анализ морфологических характеристик кожи на клеточном уровне, что существенно повышает их стоимость и усложняет техническую реализацию [1-3]. В связи с этим актуальной является задача разработки методов и технических средств, реализующих возможность проведения экспресс-диагностики заболеваний и системных нарушений кожи, а также позволяющих осуществлять контроль эффективности терапевтических процедур. Одним из перспективных направлений в изучении системы микроциркуляции является лазерная спекл-визуализация, основанная на использовании лазерного излучения для исследования биоспеклов кожи.

Актуальной задачей данного направления является разработка устройства и программного обеспечения для реализации метода исследований динамических биоспеклов. Данный метод обеспечивает визуализацию кровеносных сосудов и обнаружение в исследуемой области без инвазивного вмешательства относительных изменений капиллярного кровотока, связанных со снижением либо повышением его интенсивности. [1,5].

Биологические ткани являются оптически неоднородными поглощающими средами, средний показатель преломления которых выше, чем у воздуха, поэтому взаимодействие лазерного излучения с ними определяется процессами отражения, поглощения, рассеивания и проникновения. [7]. Метод исследования динамических биоспеклов кожи основывается на анализе параметров динамического спекл-поля, которое образуется в результате интерференции отраженного или рассеянного биообъектом когерентного излучения. Спекл-поле в плоскости наблюдения формирует картину, состоящую из множества спеклов (пятен), интенсивность света и форма которых изменяются при наличии в объекте движущихся рассеивателей

(клетки покровной ткани и форменные элементы крови) [6].

Оптические свойства дермы и скорость кровотока изменяются не только при развитии патологических процессов в организме человека (гипо- и гипертермия, посттравматическое нарушение кровоснабжения конечностей, диабетическая микроангиопатия, экзема, ангииты кожи и онкологические заболевания кожи), но и возникают как ответ на различные внешние физиотерапевтические факторы воздействия [3].

В качестве физиотерапевтических факторов активации терморегуляционных механизмов организма человека, сопровождающихся в том числе и изменениями в микроциркуляции, могут выступать различные физиотерапевтические процедуры: инфракрасная (ИК) терапия, криотерапия, гипербарическая оксигенация (ГБО).

Влияние инфракрасного излучения на организм человека проявляется в его нагреве, который способствует расширению и увеличению количества функционирующих капилляров в покровных тканях тела человека, облегчению продвижения крови по артериям, повышению скорости кровотока [8].

Криотерапия представляет собой совокупность физических методов лечения, основанных на использовании холодового фактора для отведения тепла от тканей, органов или всего тела человека, в результате чего их температура снижается в пределах криоустойчивости (5-10 С) без выраженных сдвигов терморегуляции организма. [9]. Воздействие холодом приводит к выраженным фазовым изменениям деятельности периферических сосудов, которые проявляются сначала спазмом мелких артерий и артериол, прекапиллярных сфинктеров, замедлением скорости кровотока и повышением вязкости крови. Фазовые изменения состояния сосудов кожи и подкожной клетчатки дают адекватную тренирующую нагрузку системе кровообращения [10].

ГБО заключается в лечении кислородом под давлением в медицинских бароаппаратах. Эффект применения ГБО проявляется в увеличении кислородной ёмкости крови. Сущность метода заключается в повышении содержания кислорода в тканях организма, что достигается вдыханием кислорода под повышенным давлением. Под влиянием кислородного насыщения стимулируются и нормализуются биохимические процессы в мозге, миокарде, печени. ГБО мобилизует собственные системы организма, отвечающие за обезвреживание и выведение токсинов, повышает метаболические системы защиты мозга, сердца, печени, почек от отравляющего воздействия аммиака при нарушении кровоснабжения органов.

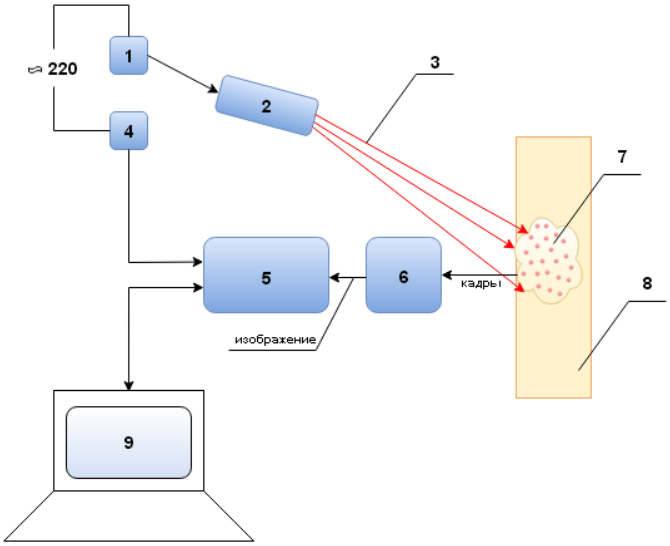
В данной работе представлены результаты применения разработанного авторами аппаратного и программного обеспечения динамического измерения биоспеклов для мониторинга микроциркуляции человека до и после проведения физиотерапевтических процедур.

Аппаратное и программное обеспечение исследований биоспеклов кожи. Регистрация динамических биоспеклов кожи выполнялась на базе устройства, схема которого приведенного на рисунке 1. На исследуемый участок кожного покрова человека фокусируется пучок лазерного излучения, интерференционная картина рассеянного биообъектом лазерного излучения регистрируется с помощью видеокамеры, снабженной специальной оптической системой. Полученная видеоинформация поступает на персональный компьютер для отображения и цифровой обработки.

Для видеорегистрации динамических спекл-полей использовалась высокоскоростная камера с интерфейсом GigE, объективом Kowa LM50HC, CCD-матрицей и частотой 120 кадров в секунду при разрешении VGA.

Регистрируемые со скоростью 120 кадров в секунду спекл-изображения подвергались цифровой обработке. Целью цифровой обработки являлся расчет контрастности для каждого пикселя спекл-изображения.

Для реализации поставленной задачи выполнялась пространственно-временная обработка спекл-изображений на базе модификации метода LASCA.



1 – блок питания лазера; 2 – лазер; 3 – лазерное излучение; 4 – блок питания видеокамеры; 5 – видеокамера; 6 – оптическая система; 7 – спекл-картина; 8 – биообъект, 9 – персональный компьютер

Рис. 1. Схема (а) и реальный вид (б) устройства динамического измерения биоспеклов кожи

В соответствии с методом tLASCA расчет значения контрастности для каждого пикселя спекл-изображения выполняется для центральной точки в окне 3×3 по n=10 накопленным кадрам:

$$K_{\mathrm{tLASCA(i,j)}} = \frac{1}{9} \cdot \sum_{r=i-1}^{r=i+1} \sum_{c=j-1}^{c=j+1} \frac{\sigma_{i,j,t}}{\langle I_{i,j,t} \rangle}, \qquad (1)$$

где $\sigma_{i,j,t}$ – среднеквадратическое отклонение всех пикселей в пространственной (i, j) и временной (t) областях, полученное для векторизированной трёхмерной матрицы;

$I_{i,j,t}$ – среднее арифметическое значение интенсивности всех пикселей в пространственной (i, j) и временной (t) областях.

Таким образом, при использовании окна минимального размера единичный пиксель результирующего кадра содержит в себе данные до 90 соседних в пространстве и времени пикселей. При разрешении камеры 659 на 494 пикселей в обработке участвуют более 38 миллионов точек ежесекундно.

Методика исследований. Диагностика состояния кожного покрова человека выполнялась на базе устройства динамического измерения биоспеклов кожи с последующей цифровой обработкой спекл-изображений [6].

Исследования проводились на базе ИК камеры для низкоинтенсивного воздействия на тело человека (рис. 2, а), представляющей собой кабину с входной дверью, воздушными отверстиями и размещенными внутри нее источниками инфракрасного излучения; барокамеры «Vitaeris 320 Hyperbaric Chamber» (рис. 2, б); криосауны «Kältekammer –110°C» (рис. 2, в).

Объектом исследования являлся участок кожи на пальце правой руки в форме квадрата 10×10 мм. Расстояния от источника света, а также от объектива камеры до исследуемого участка составляли 275 мм.

Рис. 2. Физиотерапевтические аппараты: (а) ИК камера; (б) барокамера «Vitaeris 320 Hyperbaric Chamber»; (в) криосауна «Kältekammer –110°C»

На первом этапе исследования проводились в ИК камере. Продолжительность сеанса составила 30 минут. Температура внутри кабины во время сеанса составляла 39°С. Регистрировалось исходное состояние микроциркуляции исследуемых участков кожи (0 мин), после завершения сеанса (30 мин), и спустя 30 минут после завершения сеанса (60 мин). Визуальных изменений на исследуемых участках после проведения процедуры не наблюдалось. Полученные спекл-изображения показывают, что ведущей реакцией микроциркулярного кровотока в условиях гипертермии явилась его выраженная интенсификация (рис.3).

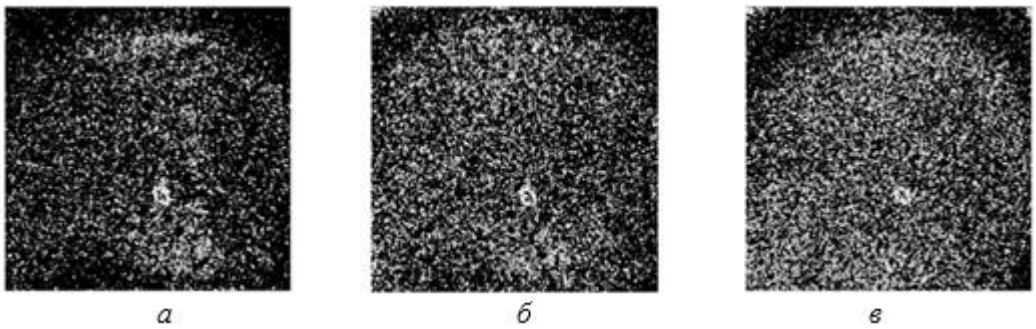


Рис.3. Спекл-изображения микроциркуляции до проведении сеанса ИК терапии (а), непосредственно после окончания сеанса ИК терапии (б), через 30 мин после после окончания сеанса ИК терапии (в)

На втором этапе исследования проводились в Барокамере «Vitaeris 320 Hyperbaric Chamber». Продолжительность сеанса составила 20 минут. Регистрация микроциркуляции проводилась в исходном состоянии (0 мин), после процедуры (20 мин), спустя 30 мин после окончания процедуры (50 мин). После процедуры гипербарической оксигенации наблюдается значительное снижение кровотока, который уже через 30 мин после окончания процедуры восстанавливается (рис.4).

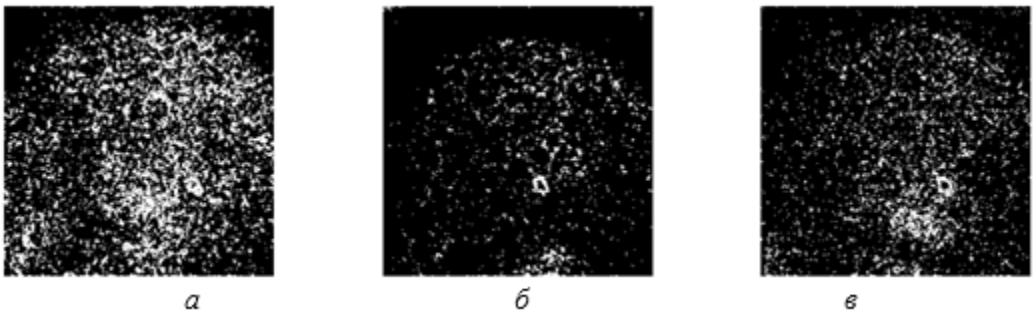


Рисунок 3 – Спекл-изображения микроциркуляции до проведении сеанса процедуры гипербарической оксигенации (а), непосредственно после окончания сеанса процедуры гипербарической оксигенации (б), через 30 мин после после окончания сеанса процедуры гипербарической оксигенации (в)

На третьем этапе исследования проводились в криосауне «Kältekammer –110°C». Испытуемый помещался в кабину на 3 минуты, температура внутри кабины составляла –110ºС. Полученные в результате обработки спекл-изображения (рис.5) отображают снижение процессов микроциркуляции в поверхностных тканях человека непосредственно после окончания процедуры криотерапии и постепенное их восстановление спустя 30 минут.

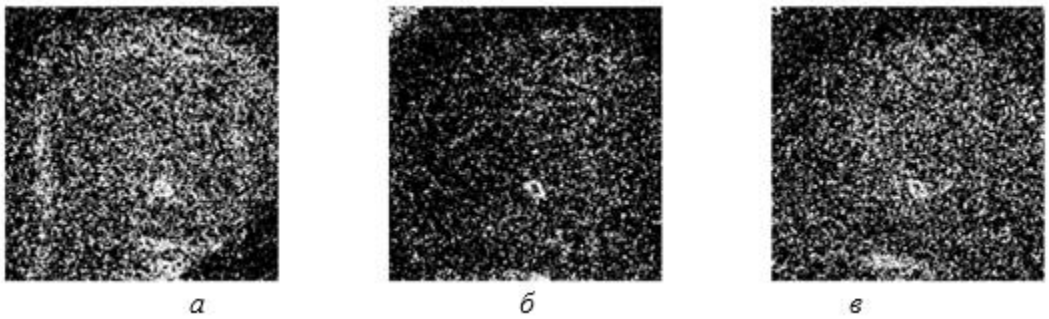


Рис. 5. Спекл-изображения микроциркуляции до проведении сеанса криотерапии (а), непосредственно после окончания сеанса криотерапии (б), через 30 мин после после окончания сеанса криотерапии (в)

Результаты исследований. Разработанное авторами аппаратное и программное обеспечение динамического измерения биоспеклов использовано для мониторинга состояния системы микроциркуляции при воздействии различных физиотерапевтических факторов.

На основании проведенных исследований были сформулированы следующие выводы:

1 Метод динамического измерения и цифровой обработки биоспеклов кожи позволяет проводить неинвазивную диагностику в режиме реального времени и получать оптические изображения внутренней структуры поверхностного кровотока.

2 Метод динамического измерения и цифровой обработки биоспеклов кожи позволяет выявить изменения в микроциркуляции, происходящие при проведении различных физиотерапевтических процедур.

3 Результаты анализа спекл-изображений согласуются с общей реакцией микроциркулярного кровотока в условиях различных воздействий на покровные ткани человека.

Мониторинг состояния системы микроциркуляции при воздействии различных физиоте-

рапевтических факторов позволяет оценить эффективность проводимых мероприятий по профилактике и лечению функциональных систем человека.

С другой стороны мониторинг состояния системы микроциркуляции содержит диагностическую информацию, так как позволяет оценить происходящие в организме естественные адаптивные процессы терморегуляции. Значения времени для возвращения уровня микроциркуляции в исходное состояние после окончания физиотерапевтической процедуры вариабельны у каждого человека, однако важным критерием нормального функционирования регуляторных механизмов является тенденция к восстановлению исходных функциональных показателей и их последующее достижение в пределах временных параметров нормы. Иная тенденция к восстановлению функциональных показателей является поводом для прекращения сеансов физиотерапевтических процедур и последующей консультации с врачом.

Результаты исследований показали целесообразность применения метода регистрации и анализа биоспеклов для оценки эффективности проводимых физиотерапевтических процедур. Кроме того, указанный метод может быть использован в разработке лечебно-диагностических комплексов, основанных на воздействии физиотерапевтических факторов, для обеспечения функций дигностики и контроля состояния микроциркуляции поверхностных биотканей, а также управления режимами воздействия.

Литература

[1]. Тимошина П.А. Мониторинг микроциркуляции крови методом спекл-контрастной визуализации в исследованиях модельных патологий на животных диссертация на соискание ученой степени кандидата физико-математических наук: 03.01.02. – Саратов, 2016. – 102 с.

[2]. Дик, С. К. Лазерно-оптические методы и технические средства контроля функционального состояния биообъектов / С. К. Дик. – Минск : БГУИР, 2014. – 235 с.

[3]. Штиршнайдер, Ю. Ю. Современные неинвазивные технологии визуализации в дерматологии / Ю. Ю. Штиршнайдер, А. В. Минченко, О. Р. Катунина, А. Р. Зубарев. – Вестник дерматологии и венерологии, вып. №5, 2011, с. 41-53.

[4]. Виленский М.А. Спекл-корреляционный анализ микрокапиллярного кровотока ногтевого ложа / М. А. Виленский, Д. Н. Агафонов, Д. А. Зимняков, В. В. Тучин, Р. А. Задражевский. – Квантовая электроника, Т.41, №4 (2011) – С.324-328.

[5]. Семячкина-Глушаковская О.В. Лазерная спекл-визуализация автономии мозгового кровообращения на уровне макро- и микроциркуляции у крыс / О.В. Семячкина-Глушаковская, А.С. Абдурашитов, С.С. Синдеев, В.В. Тучин. – Квантовая электроника, Т.46, №6 (2016) – С.496-501.

[6]. Дик С.К., Меженная М.М., Завацкий Д.А., Гордейчук Т.В., Счастная Н.И. Цифровая обработка спекл-изображений в лазерной диагностике биологических тканей Сборник материалов Второй Международной Научно-Практической Конференции «BIG DATA and Advanced Analytics BIG DATA и анализ высокого уровня» 15 — 17 июня, 2016 Минск, Беларусь. – С.282-289.

[7]. Барун, В.В., Иванов, А.П., Волотовская, А.В. / / ЖПС. 2007. Т 74. С. 391-398.

[8]. Пономаренко Г.М. Биофизические основы физиотерапии / Г.Н. Пономаренко, И.И. Турковский. М.: "Медицина", 2006. с. 17-18

[9]. Физиотерапия: национальное руководство / под ред. Г. Н. Пономаренко. – М. : ГЭОТАР-Медиа, 2009. – 864 с.

[10]. Волотовская А.В. К 82 Криотерапия: учеб.-метод. пособие / А.В. Волотовская, Г.К.Колтович, Л.Е. Козловская, А.Н. Мумин,. – Минск: БелМАПО, 2010. – 26 с

ЭТИКА БОЛЬШИХ ДАННЫХ

Д.А. Пархоменко
Старший преподаватель кафедры инженерной психологии и эргономики БГУИР, магистр техники и технологии

В.В. Шаталова
Заместитель декана по учебно-методической работе факультета компьютерного проектирования БГУИР, доцент кафедры проектирования информационно-компьютерных систем БГУИР, кандидат технических наук, доцент

Белорусский государственный университет информатики и радиоэлектроники, Республика Беларусь
E-mail: parkhomenko@bsuir.by, shatalova@bsuir.by

Анализ больших данных сегодня является не только трендом, но и необходимостью, которая может дать неоспоримое конкурентное преимущество. Технологии работы с большими данными могут принести огромную пользу как отдельным лицам, так и организациям, позволяя персонализировать обслуживание, обнаруживать мошенничество и злоупотребления, эффективное использовать ресурсы, предотвращать аварии или сбои в работе той или иной системы.

Одновременно с развитием технологии анализа больших данных встает вопрос об этике такой аналитики. По своей сути данная технология является этико-агностической, но это подталкивает к преодолению различных ограничений: наличия широкого диапазона данных из многих источников; возможность недорогостоящей корреляции этих данных для понимания более широкой картины; точность, с которой человек может быть идентифицирован и подвергнут таргетированию; умение определять местонахождение кого-либо для контекстного анализа и наблюдения; применение этого нового понимания для широкого круга решений и действий; работа в режиме реального времени.

Сегодня законы и нормативные документы не успевают отвечать тому, что порождают новые цифровые технологии. Конфиденциальность личных данных под угрозой. Также пользователям не хватает знаний и понимания таких технологий, чтобы защитить себя. Недавние достижения в аналитике больших данных расширили разрыв между тем, что возможно и что разрешено законом, изменении баланс сил между отдельными лицами и агрегаторами данных. Сегодня появляются новые возможности, связанные с социальными катастрофами и непредвиденными последствиями. Именно в этом промежутке поднимаются этические вопросы вокруг того, что приемлемо. Как организация применяет аналитику больших данных для повышения эффективности своей работы, понимает что использование этой технологии является этичным? Кто решает, как и когда правильно пользоваться большими данными? Согласно прогнозам аналитического агентства «Gartner», к 2018 году около половины всех нарушений деловой этики будет обусловлена неправильным использованием имеющихся средств анализа больших данных.

Примеры хорошей и плохой практики появляются в отрасли, и со временем они будут определять правила и законодательство.

История сети супермаркетов «Target», о том что алгоритм определил беременность девочки и стал присылать ей купоны на товары для беременных, еще до того как семья девочки узнала о ее беременности, часто звучит как пример возможностей больших данных и машинного обучения. Однако эта история подняла волну протеста, и «Target» потеряла лояльность своих клиентов.

Проблемы, которыми занимается цифровая этика, можно объединить в несколько групп:

– проблема «Выбора» – когда вам нужно выбрать один из нескольких вариантов, каждый из которых является приемлемым для вас и не касается вас лично;

– проблема «Собственника» – когда вы должны сделать личный выбор – вы хотите заработать много денег или сделать вашего клиента счастливым;

– проблема «Времени» – когда вы должны выбрать между тактическим и стратегическим решением.

Профессор права Вашингтонского университета Нил Ричардс отмечает: «этика диктует принципы, которыми руководствуются законодатели». Принятие законов, касающихся цифровой сферы сильно запаздывает. Сегодня закон практически не регламентирует манипуляции с большими данными, хотя уже возникают моменты, которые люди воспринимают с возмущением.

Для того, чтобы компаниям избежать потери лояльности клиентов в новых рыночных условиях использования аналитики больших данных, компаниям необходимо вести себя прозрачно и обсуждать вопросы использования данных непосредственно с потребителями.

ПРИМЕНЕНИЕ ТЕХНОЛОГИЙ BIG DATA ДЛЯ ПОСТРОЕНИЯ АНТРОПОМОРФНОЙ СИСТЕМЫ УПРАВЛЕНИЯ ПРОМЫШЛЕННЫМИ РОБОТЕХНИЧЕСКИМИ КОМПЛЕКСАМИ

М.В. Давыдов
Заведующий кафедрой теоретических основ электротехники БГУИР, кандидат технических наук, доцент

Н.С. Давыдова
Доцент кафедры сетей и устройств телекоммуникаций БГУИР

А.Н. Осипов
Проректор по научной работе БГУИР, кандидат технических наук, доцент

Белорусский государственный университет информатики и радиоэлектроники, Республика Беларусь
E-mail: davydov-mv@bsuir.by

Abstract. В статье предложена антропоморфная система управления промышленными роботехническими комплексами. Данная система включает 5 уровней управления и обладает рядом преимуществ по сравнению с классическими системами управления техническими объектами. Показано, что технологии big data могут применятся для обработки многочисленных информационных сигналов обратной связи, необходимых для организации управления промышленными роботехническими комплексами.

Одним из основных направлений развития технических систем в настоящее время является создание роботов и роботизированных комплексов [1-2]. Области их применения постоянно расширяются: космос [3], военное, и промышленное применение [4–8], медицина [9-10], социальная работа, системы безопасности [11], умный дом и др. При этом современные роботизированные комплексы постоянно усложняются, например, в промышленности с введением стандарта промышленного интернета вещей [12-13] есть тенденция объединения отдельных роботов в единые интеллектуальные роботизированные производственные линии. Все это приводит к существенному усложнению аппаратных и программных средств, предназначенных для управления роботом или роботизированным комплексом. Разработке аппаратных и программных моделей систем управления роботами посвящен ряд работ [14 – 18]. В работе [19] авторами предложена модель системы управления, построенная по аналогии с деятельностью центральной нервной системы человека. В данной работе рассмотрены аспекты применения технологий big data для обработки сигналов обратной связи, необходимых для организации управления промышленными роботехническими комплексами.

На основе созданной структурной схемы, используя антропоморфный принцип проектирования иерархических технических систем [20, 21], разработана многоуровневая структурная схема управления роботехническим комплексом (рисунок 1).

Данная схема включает следующие уровни управления:

1. Уровень А – Драйверы исполнительных устройств. Данный уровень обеспечивает управление конечным исполнительным устройством, например шаговым или асинхронным двигателем, сервоприводом, пневмо- или гидроприводом и т.д. Драйверы могут иметь различные информационные входы управления – аналоговый сигнал тока или напряжения, цифровой код, сложный цифровой протокол. Как правило на драйвер поступает диагностическая информация от исполнительного устройства: потребление тока, сигнал энкодера, наличие перемещения и т.д. Для обеспечения высокой точности движений драйверы могут также учитывать

пространственную информацию – сигнал с датчиков положения, акселерометров, гироскопов. Драйвер также формирует информационный сигнал, содержащий диагностическую информацию о текущем состоянии исполнительного устройства.

2. Уровень B – Контроллер формирования команд. Данный контроллер представляет собой блок управления, к которому посредством драйверов подключена группа исполнительных устройств, для которых контроллер генерирует требуемые управляющие сигналы. Кроме того в памяти данного блока уже хранятся простейшие двигательные паттерны для групп исполнительных устройств.

Входными данными для блока является 1) информация о требуемом в текущий момент времени движении (уровень C); 2) информация с датчиков положения, акселерометров, гироскопов и т.д. позволяющая осуществлять контроль выполнения сложного движения обеспечиваемого группой исполнительных устройств; 3) информация о состоянии конечных исполнительных устройств с блока диагностики состояния поступает в контроллер и учитывается при формировании управляющих воздействий.

Техническая реализация данного уровня управления возможна на базе высокопроизводительных микроконтроллеров (например STM32 ARM Cortex), либо с помощью программируемых логических контроллеров.

3. Уровень C – Подсистема анализа окружающей обстановки и формирования алгоритма решения текущей задачи. Данная система предназначена для построения алгоритма работы всех исполнительных устройств при решении текущей задачи. Входными данными для подсистемы является информация от подсистемы анализа целей (уровень D). Подсистема учитывает информацию о положении и перемещении манипуляторов, их состояние (информация с уровня B), текущее положение системы в пространстве относительно других предметов (информация с оптических и ультразвуковых датчиков, а также концевых переключателей) и формирует общий алгоритм действия и выдает информацию о требуемых в текущий момент движениях на уровень B.

4. Уровень D – Подсистема анализа целей (семантическая система). Входными данными этой системы являются общая, в некоторых случаях неформализованная задача, поступающая от оператора либо другой технической системы посредством интерфейса человек-машина либо машина-машина. Данная подсистема служит для общего анализа поставленной задачи, декомпозиции и построения общего плана действий для ее решения. При этом должны учитываться техническое состояние исполнительных систем (информация с блока диагностики состояния технических систем) а также текущее положение и состояние окружающей обстановки (информация с уровня C).

5. Уровень E – Интерфейс человек-машина/машина-машина. Данный уровень служит для преобразования команд оператора либо другой технической системы в формат «понятный» семантической системе.

Как правило реализация уровней C, D и E производится программными средствами. В простейшем случае их можно реализовать на базе одноплатных компьютеров (Raspberry Pi, Red Pitaya, ZedBoard) для более совершенных систем могут использоваться более производительные промышленные компьютеры.

Как видно из представленного описания данная система управления отличается большим количеством информационных связей и к данным поступающим в блок управления можно применить классические отличительные признаки big data: объем, скорость и разнообразие (*VVV*: *volume*, *velocity*, *variety*).

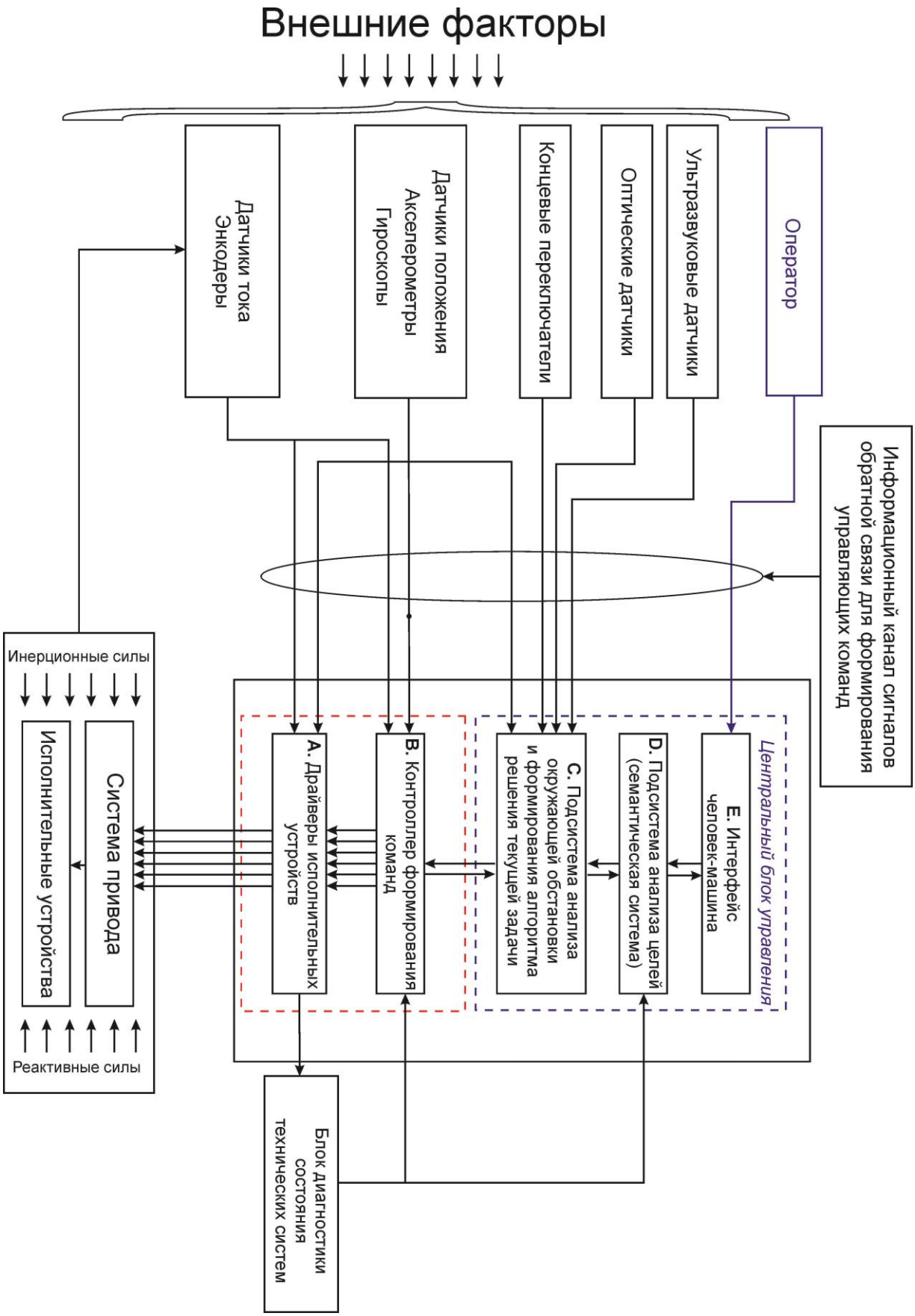


Рис. 1. Многоуровневая структурная схема управления роботехническим комплексом

Разработанные технологии для работы с big data пока не отвечают требованиям встроенных систем. Однако обращение к классической форме обработки big data [22, 23] позволяет предложить подход для кластеризации и предобработки получаемых данных (рисунок 2).

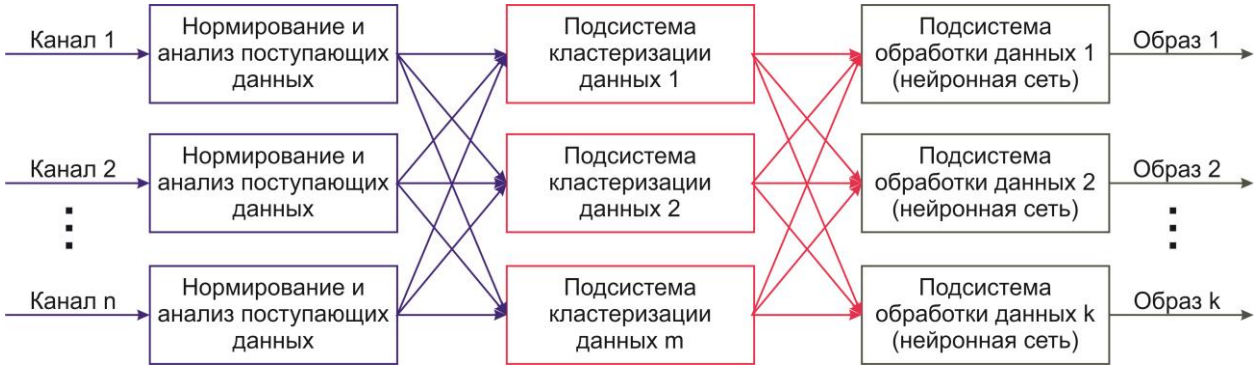


Рис. 2. Модель предобработки информационных сигналов обратной связи

Информация от датчиков должна проходить три этапа:

1. происходит нормирование и анализ информации в каждом информационном канале (1… n);
2. выполняется кластеризация данных для последующей обработки (количество кластеров определяется логической структурой получаемых данных: позиция, положение, параметр, команда и т.д. 1… m);
3. происходит обработка и передача обобщенных данных (образов) на соответствующие уровни системы управления (1…k).

Предложенная структура управления обладает рядом достоинств:

1 После предобработки данные в виде образов поступают на соответствующие уровни управления, что упрощает их интерпретацию. При добавлении информации соответствующей коррекции требует только система предобработки, но не сама система управления.

2 Уровни иерархии системы управления хорошо разграничивают функции управления. Таким образом, разработку подобной системы легко распараллелить.

3 В последствии, достаточно просто выполнять модернизацию системы путем модернизации каждого уровня управления по-отдельности.

4 Данная система управления может применяться не только для управления одним роботехническим комплексом, она может масштабироваться с целью управления группой роботехнических комплексов или целым роботизированным предприятием (при введении шестого технологического уклада).

Авторы выражают признательность доктору технических наук, профессору Голенкову Владимиру Васильевичу за оказанную помощь при обсуждении и написании настоящей статьи.

Литература

[1]. Отчет ЦЭМИ РАН от 2015 г. «Революционные технологии: перспективные направления развития робототехники» Программа Президиума РАН «Анализ и прогноз долгосрочных тенденций научного и технологического развития: Россия и мир».

[2]. Комков Н.И., Бондарева Н.Н. Перспективы и условия развития робототехники в России // МИР (Модернизация. Инновации. Развитие). 2016. № 2.

[3]. Huang P. et al. Dynamics and configuration control of the maneuvering-net space robot system //Advances in Space Research. – 2015. – Т. 55. – №. 4. – С. 1004-1014.

[4]. Дульнев П. А. К вопросу о роботизации вооружения и военной техники сухопутных войск //Вестник академии военных наук. -- 2015 – №. 1(50),

[5]. Рубцов И. В. Вопросы состояния и перспективы развития отечественной наземной робототехники военного и специального назначения //Известия Южного федерального университета. Технические науки. – 2013. – №. 3 (140).

[6]. Kuss A. et al. Manufacturing knowledge for industrial robot systems: Review and synthesis of model architecture //Automation Science and Engineering (CASE), 2016 IEEE International Conference on. – IEEE, 2016. – С. 348-354.

[7]. Kaltsoukalas K., Makris S., Chryssolouris G. On generating the motion of industrial robot manipulators //Robotics and Computer-Integrated Manufacturing. – 2015. – Т. 32. – С. 65-71.

[8]. Li J. et al. A design pattern for industrial robot: user-customized configuration engineering //Robotics and Computer-Integrated Manufacturing. – 2015. – Т. 31. – С. 30-39.

[9]. Kraus P., Geiger R. Robot system for medical surgeries : пат. D768219 США. – 2016.

[10]. Joubair A. et al. Absolute accuracy analysis and improvement of a hybrid 6-DOF medical robot //Industrial Robot: An International Journal. – 2015. – Т. 42. – №. 1. – С. 44-53.

[11]. Liu J. N. K., Wang M., Feng B. iBotGuard: an Internet-based intelligent robot security system using invariant face recognition against intruder //IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews). – 2005. – Т. 35. – №. 1. – С. 97-105.

[12]. Atzori L., Iera A., Morabito G. The internet of things: A survey //Computer networks. – 2010. – Т. 54. – №. 15. – С. 2787-2805.

[13]. Sadeghi A. R., Wachsmann C., Waidner M. Security and privacy challenges in industrial internet of things //Proceedings of the 52nd Annual Design Automation Conference. – ACM, 2015. – С. 54.

[14]. Евграфов, В.В. Динамика, управление, моделирование роботов с дифференциальным приводом Текст. /В.В. Евграфов, В.Е. Павловский, В.В. Павловский // Теория и системы управления. 2007. - №5. - С. 171-176.

[15]. Глазкова Л. В., Панченко А. В., Павловский В. Е. Динамика, моделирование и управление колёсным робобуером // Нелинейная динамика. — 2012. — Т. 8, № 4. — С. 679—687

[16]. Павловский В.Е., Павловский В.В. Модульная микроконтроллерная система управления роботами РОБОКОН-1 // Препринты ИПМ им. М.В.Келдыша. 2012. № 86. 32 с.

[17]. Прокопович Г. А. Нейросетевая модель для реализации поисковых движений мобильного робота. – 2013.

[18]. Прокопович, Г.А. Способ управления манипулятором робота на основе гетеро-ассоциативной искусственной нейронной сети / Г.А. Прокопович // Электроника Инфо (рецензируемый раздел). – 2014. - №6 (108). – С. 36-39.

[19]. Давыдов, М.В. Антропоморфная система управления роботехническим комплексом/ Давыдов М.В., Давыдова Н.С., Осипов А.Н.// Открытые семантические технологии проектирования интеллектуальных систем = Open Semantic Technologies for Intelligent Systems : материалы междунар. науч.-техн. конф./ редкол. : В. В. Голенков (отв. ред.) [и др.], ISSN 2415-7740; Вып.1 (Минск, 16-18 февраля 2017г.). – Минск : БГУИР, 2017. -- 466 с.

[20]. Бажин С. А., Васильевский А. С., Лапшин К. В. Стратегия проектирования антропоморфных систем //Информационно-управляющие системы. – 2012. – №. 5 (60).

[21]. Мако Д., Месарович М., Такахара И. Теория иерархических многоуровневых систем //Мир. – 1973.

[22]. Zikopoulos P. et al. Understanding big data: Analytics for enterprise class hadoop and streaming data. – McGraw-Hill Osborne Media, 2011.

[23]. Sharma A. B. et al. Modeling and analytics for cyber-physical systems in the age of big data //ACM SIGMETRICS Performance Evaluation Review. – 2014. – Т. 41. – №. 4. – С. 74-77.

AUTOR INDEX

СПОНСОР КОНФЕРЕНЦИИ

itransition
SOFTWARE DEVELOPMENT COMPANY

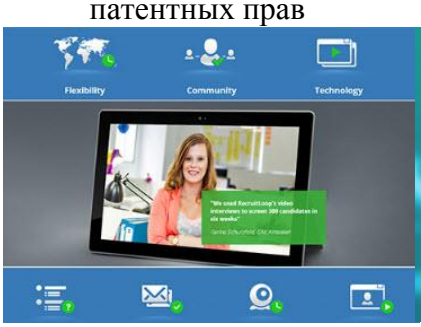
Международная ИТ-компания с 15-летней историей, предоставляющая весь комплекс услуг в сфере разработки ПО.

Отслеживание транспортных перевозок

Patriot: онлайн-платформа для выявления нарушений патентных прав

Интерактивный образовательный портал

Образовательный портал для PayPal Inc.



Подбор персонала онлайн

Комплексная система бизнес-аналитики

Система управления документами для оператора связи

WHEN YOU WISH
Платформа по сбору средств

Экосистема Facebook-игр, управляемая единой платформой

СПОНСОР КОНФЕРЕНЦИИ

ALTOROS™

Разработка софтверных решений.
Интеграция Cloud Foundy с SQL, NoSQL, Hadoop

Multi-cloud Deployment Automation

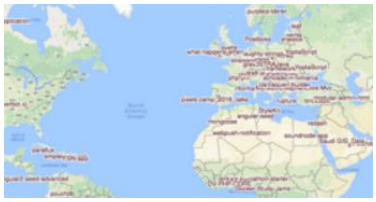
Integration of Cloud Foundry with SQL / NoSQL / Hadoop

Cloud Foundry Deployment

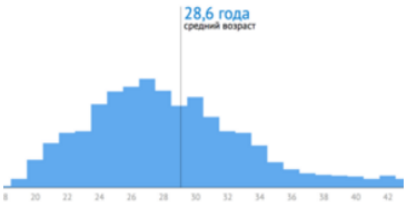
СПОНСОР КОНФЕРЕНЦИИ

dev.by

Интернет-ресурс про работу в IT



YoptaScript и ClickHouse: топ-10 самых популярных в Беларуси проектов на GitHub



ИТ в Беларуси-2016: в индустрии ещё никогда не было столько новичков

Не вместо, а вместе: автоматизация увеличивает количество рабочих мест?

СПОНСОР КОНФЕРЕНЦИИ

Компания занимает лидерскую позицию, войдя в список Global Services 100 и другие крупнейшие в своей отрасли мировые рейтинги

РАЗРАБОТКА ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ

МОБИЛЬНОЕ И ВСТРОЕННОЕ ПО

РЕШЕНИЯ ДЛЯ ЭЛЕКТРОННОЙ КОММЕРЦИИ

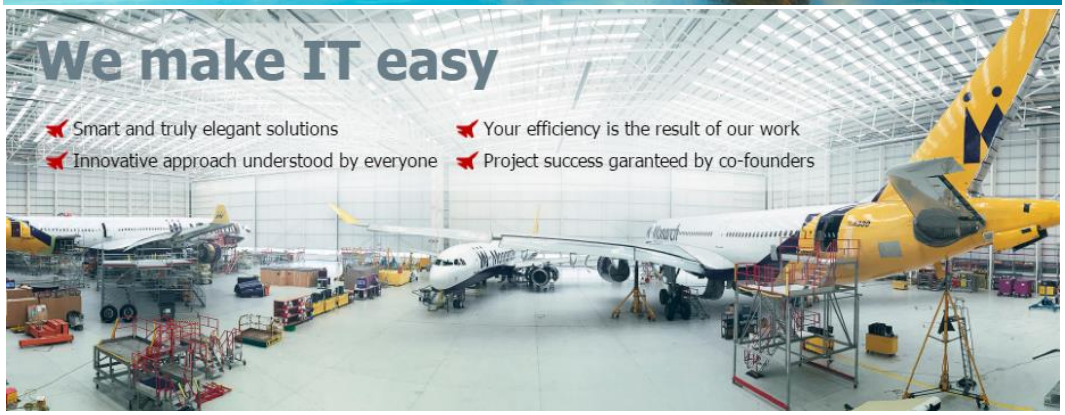
ПОСТРОЕНИЕ ВЫДЕЛЕННЫХ КОМАНД

СПОНСОР КОНФЕРЕНЦИИ

Специализация компании - разработка программного обеспечения и оказание консультационных услуг в сфере разработки ПО. С 2005 года Artezio входит в состав группы компаний ЛАНИТ. Специалистами Artezio выполнено более 1000 проектов для клиентов из России, Западной Европы, Израиля, Японии, США и Канады.

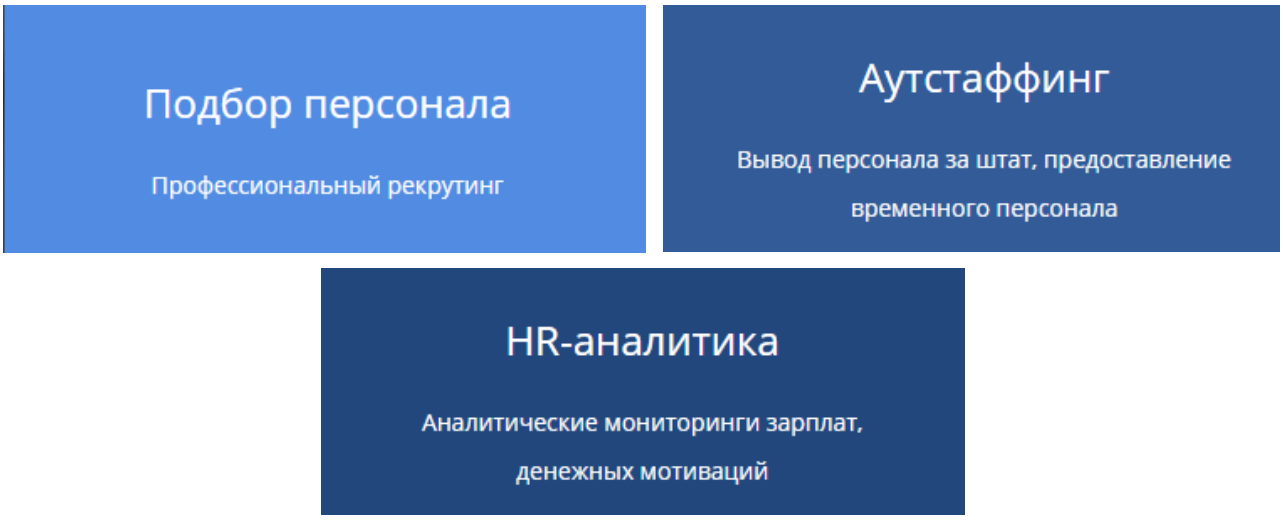
СПОНСОР КОНФЕРЕНЦИИ

ООО "Джет Би Ай" создано в декабре 2013 года и специализируется на разработке корпоративного программного обеспечения для бизнес-аналитики и автоматизации процессов стратегического маркетинга на базе технических инструментов и программных платформ SAP и Salesforce.



СПОНСОР КОНФЕРЕНЦИИ

Группа компаний KIAT позиционирует себя как HR- провайдера, предоставляющего клиентам линейку высококачественных HR- сервисов. «КИАТ» начал работать в 2002 году как классическое кадровое агентство



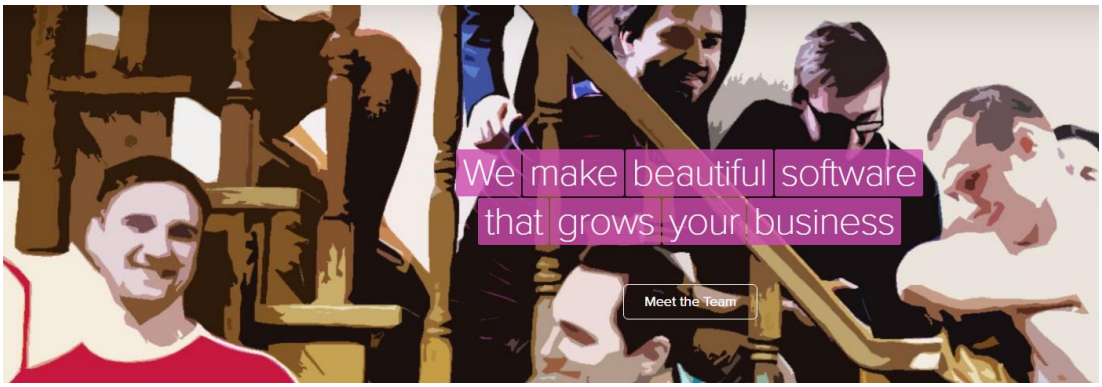
СПОНСОР КОНФЕРЕНЦИИ

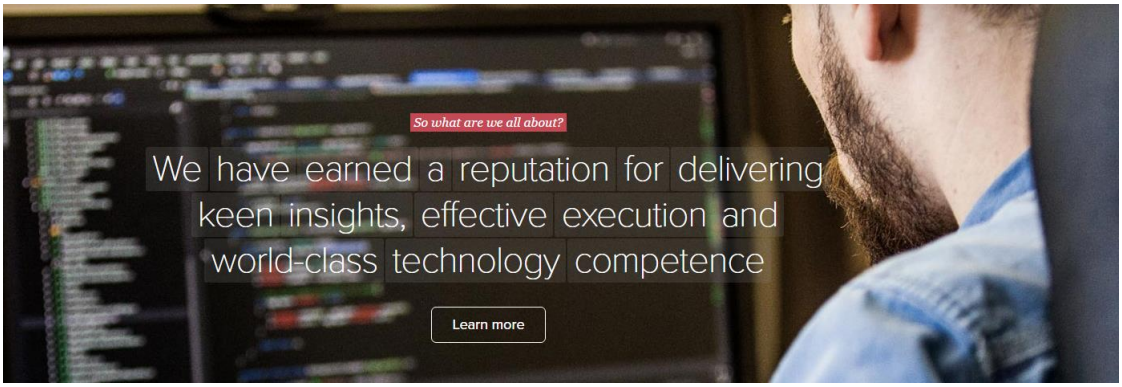
ООО «Лаборатория цифровых систем» занимается разработкой интеллектуального программного обеспечения и приложений для анализа данных, применяя технологии машинного обучения.

СПОНСОР КОНФЕРЕНЦИИ

Paralect

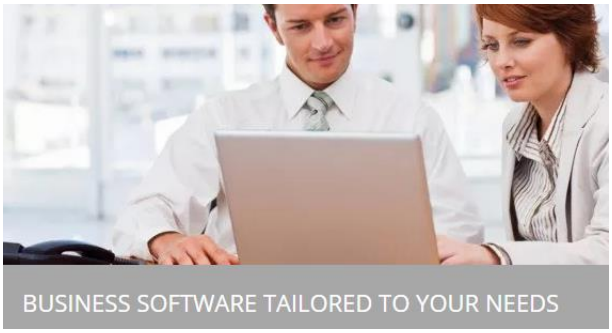
Paralect создает технологические продукты и услуги для клиентов по всему миру. Небольшая и дружественная компания технологического программного обеспечения в Минске, которая была разработана с нуля, чтобы быть отличным местом для работы.

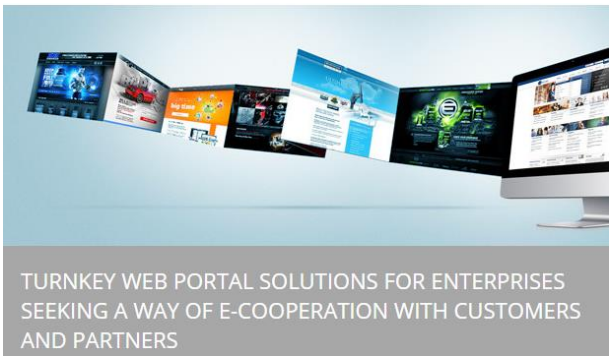




СПОНСОР КОНФЕРЕНЦИИ

ScienceSoft Inc. (ЗАО «НАУЧСОФТ») – белорусская компания-разработчик программного обеспечения. Услуги компании включают разработку программных продуктов и решений, разработку приложений для мобильных устройств, ИТ-консалтинг, услуги по тестированию, разработке баз данных и услуги по работе с IT-инфраструктурой.





СПОНСОР КОНФЕРЕНЦИИ

ИП Тимохов

Консультации в области разработки ПО. Бизнес-аналитик. IT-партнер крупного российского авиаперевозчика.

СПОНСОР КОНФЕРЕНЦИИ

GODEL TECHNOLOGIES

Разработка высокотехнологичных ИТ-решений: разработка масштабного ПО под заказ, управление циклом разработки, тестирование ПО, аутсорсинг, совместная разработка, бизнес-аналитика, сервис-ориентированная архитектура, управление идентификацией, системы управления контентом, системы управления взаимодействием с клиентами.

СПОНСОР КОНФЕРЕНЦИИ

МОЛОДЁЖНОЕ ТЕЛЕВИДЕНИЕ БГУИР

Молодежное телевидение БГУИР

СПОНСОР КОНФЕРЕНЦИИ

ООО «Эктив Технолоджис» является разработчиком программного обеспечения для автоматизации бизнес-процессов провайдеров облачных сервисов. Компания предоставляет решения для построения публичных облаков, биллинга и управления различими сервисами: SaaS (Software as a Service), PaaS (Platform as a Service), IaaS (Infrastructure as a Service), а также консалтинг в данной сфере.

СПОНСОР КОНФЕРЕНЦИИ

ООО «Нэкстсофт» занимается разработкой программного обеспечения системной инфраструктуры; программного обеспечения, предназначенного для разработки и развертывания прикладных программ; ИТ-аутсорсингом.

СПОНСОР КОНФЕРЕНЦИИ

СПОНСОР КОНФЕРЕНЦИИ

ООО «Бримит» – это веб-разработчик, специализирующийся на веб-платформе Microsoft .NET и мобильных приложениях. С помощью новейших технологий .Net и новейших интернет-рекомендаций ООО «Бримит» разрабатывает веб-решения наилучшим образом, наиболее эффективным и инновационным способом.

СПОНСОР КОНФЕРЕНЦИИ

ООО "Совершенные системы" является официальным поставщиком в Республику Беларусь оборудования и программного обеспечения торговых марок TRASSIR, ACTIVECAM и "СФИНКС".