Advanced Research in Mathematics and Computer Science

Citation for published version (APA):

Sloep, P., Stefanov, K., Zlateva, N., Koytchev, I., & Boytchev, P. (Eds.) (2014). *Advanced Research in Mathematics and Computer Science: Doctoral Conference in Mathematics, Informatics and Education [MIE 2013] Sofia, Bulgaria, 2013, Proceedings*. St. Kliment Ohridski University Press.

Document status and date:
Published: 01/09/2014

Document Version:
Publisher's PDF, also known as Version of record

Please check the document version of this publication:

• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

https://www.ou.nl/taverne-agreement

Take down policy
If you believe that this document breaches copyright please contact us at:

pure-support@ou.nl

providing details and we will investigate your claim.

Downloaded from https://research.ou.nl/ on date: 12 Oct. 2022

Open Universiteit
www.ou.nl

Advanced Research in Mathematics and Computer Science

ST. KLIMENT OHRIDSKI UNIVERSITY PRESS

Advanced Research in Mathematics and Computer Science

Doctoral Conference in Mathematics,
Informatics and Education [MIE 2013]
Sofia, Bulgaria, 2013, Proceedings

St. Kliment Ohridski University Press
Sofia • 2013

© 2013 Peter Sloup, Krassen Stefanov, Nadya Zlateva, Ivan Koytchev, Pavel Boytchev
© 2013 St. Kliment Ohridski University Press

ISBN 978-954-07-3596-2

Preface

The conference proceedings are result from the doctoral conference MIE 2013: Doctoral Conference in Mathematics, Informatics and Education, held in September 19–21, 2013, at Sofia, Bulgaria. This conference was organised as part of the activities in the project under the Human resources development scheme in Bulgaria, aiming to support doctoral students, post doctoral lecturers and young scientists.

Thirty three research papers were submitted to this international doctoral conference and 22 were accepted for final publication.

MIE 2013 is a doctoral student conference for researchers in Mathematics and Computer Science to connect with international research communities for the worldwide dissemination and sharing of research ideas and results.

Three coherently interrelated tracks were arranged in the three-day conference including Mathematics, Informatics, and Technology enhanced learning. Young researchers and post-doctoral students participated in paper presentations, doctoral student consortia and panel discussions under the themes of the conference tracks.

More information about the conference is available on the conference web site: `http://mie.uni-sofia.bg/`

The papers in the proceedings are organised in three sections, according to the corresponding track.

September 2013

Peter Sloep
Program Committee Chair
MIE2013

Organization

MIE2013 is organized by the Faculty of Mathematics and Informatics, Sofia University “St. Kliment Ohridski”.

Organizing Committee

Chair:

Krassen Stefanov

Members:

- Atanas Georgiev
- Eugenia Kovatcheva
- Eliza Stefanova
- Stanimira Yordanova
- Temenujka Zafirova-Maltcheva

Program Chair

Peter Sloep, Open University, The Netherlands

Referred Papers Track Chairs

- Nadya Zlateva, Sofia University, Bulgaria (Mathematics)
- Ivan Koychev, Sofia University, Bulgaria (Informatics)
- Pavel Boytchev, Sofia University, Bulgaria (Technology Enhanced Learning)

Program Committee

- Galia Angelova, Bulgarian Academy of Science, Bulgaria
- Marusia Bojkova, Sofia University, Bulgaria
- Boyan Bontchev, Sofia University, Bulgaria
- Tatiana Chernogorova, Sofia University, Bulgaria
- Pando Georgiev, University of Florida, USA
- Vassil Georgiev, Sofia University, Bulgaria
- Olga Georieva, Sofia University, Bulgaria
- Alexander Grigorov, Sofia University, Bulgaria
- Elissaveta Gourova, Sofia University, Bulgaria
- Christo Dichev, Winston-Salem State University, USA
- Vladimir Dimitrov, Sofia University, Bulgaria
- Milena Dobreva, University of Malta, Malta

- Gabriella Dodero, Free University of Bozen-Bolzano, Italy
- Asen Dontchev, University of Michigan, USA
- Hristo Gantchev, Sofia University, Bulgaria
- Valentin Kisimov, University of National and World Economy, Bulgaria
- Goran Velinov, Ss Cyril and Methodius University Skopje, Macedonia
- Piet Kommers, University of Twente, Netherlands
- Rob Koper, Open University, Netherlands
- Sylvia Ilieva, Sofia University, Bulgaria
- Stefan Ivanov, Sofia University, Bulgaria
- Jordan Jordanov, Sofia University, Bulgaria
- Kalinka Kaloyanova, Sofia University, Bulgaria
- Evgenii Krastev, Sofia University, Bulgaria
- Plamen Mateev, Sofia University, Bulgaria
- Leda Minkova, Sofia University, Bulgaria
- Preslav Nakov, Bulgarian Academy of Science, Bulgaria
- Geno Nikolov, Sofia University, Bulgaria
- Maria Nisheva, Sofia University, Bulgaria
- Radoslav Pavlov, Bulgarian Academy of Science, Bulgaria
- Nedyu Popivanov, Sofia University, Bulgaria
- Atanas Radenski, Chapman University, USA
- Petko Ruskov, Sofia University, Bulgaria
- Julian Revalski, Institute of Mathematics and Informatics, Bulgarian Academy of Science, Bulgaria
- Dimitar Skordev, Sofia University, Bulgaria
- Slavi Stoyanov, Open University, Netherlands
- Tinko Tinchev, Sofia University, Bulgaria
- Julita Vasileva, University of Saskatchewan, Canada
- Wim Westera, Open University, Netherlands
- Vladimir Zanev, Columbus State University, USA

Table of Contents

Mathematics

Informatics

Technology Enhanced Learning

Crump–Mode–Jagers Branching Processes: Application in Population Projections

Plamen Trayanov[1] and Maroussia Slavtchova–Bojkova[1,2]

[1] Faculty of Mathematics and Informatics, Sofia University “St. Kliment Ohridski”,
5, James Bourchier Blvd., 1164 Sofia, Bulgaria
plamentrayanov@gmail.com, bojkova@fmi.uni-sofia.bg
[2] Institute of Mathematics and Informatics, Bulgarian Academy of Sciences
Acad. G. Bonchev Str., Bl. 8, 1113 Sofia, Bulgaria

Abstract. In this paper the Crump–Mode–Jagers branching process theory is used to model human population and give projections of how it will evolve in future. How the population grows according to given scenarios can be useful in decision making and choosing an appropriate demographic policy. If we know how a social policy will affect the birth process for example, then we can estimate what will be the influence of that policy on the population growth. For that reason we suggest to use the Malthusian parameter in order to decide what kind of policy would be most relevant towards stimulating the population growth. On the other hand, by modelling the past, we can learn from it and adjust. That is why modelling population with General Branching Process (GBP) can be very useful and important.

Keywords: general branching process, demography, population projections, Malthusian parameter

Introduction

How could we predict the future demographic state and are there a proper parameters knowing which could help us to make accurate prediction? Our aim is to give answer to that question mathematically, by means of so-called Crump–Mode–Jagers branching process (CMJBP) or general branching process (GBP) (see Jagers [1]) to model a human population. Then we apply this model to make projections of Bulgarian population. It is regularly a widespreading topic in mass media that Bulgarian population is going to diminish and according to the National Statistical Institute (see [2]) the female count is expected to reach 2,697,991 in 2060. With this article we would like to shed light as to this matter by means of stochastic modelling.

The article consists of two sections. The General Branching Process (GBP) as defined in Jagers [1] as well as its particular case given in [3] are presented in Section 1. An application of the model in forecasting the future population count and some conclusions are presented in Section 2.

1 Theoretical Framework

1.1 Some Preliminaries of GBP

For the sake of clarity in this section we will briefly remind some preliminary results concerning the theory of GBP (see Jagers [1] and Trayanov [3]). Let I denote the set of all n-tuples for all values of $n \geq 1$. It is assumed the GBP starts from a single individual denoted by (0) and the n^{th} daughter of mother x is denoted by $(x, n) \in I$. A random variable (r.v.) λ_x that describes the life-length of that individual and a point process ξ_x that describes the birth process of the mother through her life are defined for each individual x. A point process ξ_x on R^+ is a map from a probability space into the set of integer or infinite valued measures on R^+ defined on the Borel σ-algebra, such that the mass $\xi(A)$ given to a bounded Borel set A is a finite r.v. (see Jagers [1]). We say the individual $(x, n) \in I$ is realized if $\xi_x(\infty) \geq n$. But, even if the set I is infinite, the number of realized individuals is finite. Thus, for each $x \in I$ a couple (λ_x, ξ_x) is defined and these couples are assumed to be identically and independently distributed. An important implication of this assumption is that actually their distributions are assumed to remain constant in time. It will be shown later that this restriction can be avoided quite simply in order to describe more accurately the reality. Let us denote by the r.v. $\tau_x(k) = \inf\{t : \xi_x(t) \geq k\}$ the mother's age at birth of child (x, k). Then $\sigma_x = \tau_0(j_1) + \tau_{j_1}(j_2) + \cdots + \tau_{(j_1,\ldots,j_{n-1})}(j_n)$, where $x = (j_1, \ldots, j_n)$ denotes the birth date of x. The first individual has birth date 0. If n^{th} child was not born at all, then his(her) birth date is infinity. An indicator variable $z_t^a(x)$ is defined for the individual x to be alive and younger than age $a > 0$ at time $t > 0$ as follows

$$z_t^a(x) = \begin{cases} 1, & t - a < \sigma_x \leq t < \sigma_x + \lambda_x, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

Definition 1. *The GBP is a stochastic process z_t^a defined as follows:*

$$z_t^a = \sum_{x \in I} z_t^a(x). \tag{2}$$

For $a > t$ we denote by $z_t = z_t^a$ the total number of individuals in the population.

Definition 2. *The function $\int_0^\infty e^{-sx}\, d\mu(x)$ is called Laplace–Stieltjes transform of the function $\mu(t)$, $t > 0$.*

Let us denote by $f(s) = \mathbb{E}\left(s^{\xi(\infty)}\right)$, $|s| \leq 1$, $L(t) = \mathbb{P}(\lambda_x \leq t)$, $\hat{\mu}$ being the Laplace–Stieltjes transform of $\mu(t) = \int_0^t f_u'(1) L(du)$ and $S(t) = 1 - L(t)$. An important result in Jagers [1, Ch. 6.3, Theorem 6.3.3] gives us a renewal equation for the expected future population count:

Theorem 1. *If* $f(s) < \infty$, $|s| \leq 1$, *then* $m_t = \mathbb{E}(z_t) < \infty$ *for all* t *and* $m_t^a = \mathbb{E}(z_t^a)$ *satisfies*

$$m_t^a = 1_{[0,a)}(t)\{1 - L(t)\} + \int_0^t m_{t-u}^a \, \mu(du), \tag{3}$$

where

$$1_{[0,a)}(t) = \begin{cases} 1, & 0 \leq t \leq a, \\ 0, & \textit{otherwise}. \end{cases}$$

If $m = \mu(\infty) < 1$, *then* $\lim\limits_{t\to\infty} m_t = 0$. *If* $m = 1$ *and* μ *is non-lattice, then for* $0 \leq a \leq x$

$$m_t^a \to \int_0^a \{1 - L(u)\} \, du \Big/ \int_0^\infty u \, \mu(du) \ .$$

If further $\int_0^\infty tL(dt) < \infty$, *then*

$$m_t^a \to \int\limits_0^\infty uL(du) \Big/ \int_0^\infty u \, \mu(du) \ .$$

If $m > 1$, μ *is non-lattice and* $\alpha > 0$ *is the Malthusian parameter defined by* $\hat{\mu}(\alpha) = 1$, *then for* $0 \leq a \leq \infty$

$$m_t^a \sim e^{\alpha t} \int_0^a e^{-\alpha u}\{1 - L(u)\} \, du \Big/ \int_0^\infty ue^{-\alpha u} \, \mu(du) \ .$$

In the lattice cases corresponding assertions hold.

Remark 1. This theorem tells us that if we know how the population reproduces (i.e. the point process $\xi(t)$) and how many years the individuals live (i.e. the r.v. λ), then we know what to expect for their count. Another useful information is the asymptotic behaviour of the population towards infinity – does it decline to 0, rise to infinity or oscillate towards a fixed number.

Another important result from Jagers [1, Ch. 6.7, Theorem 6.7.1] is as follows:

Theorem 2. *In a non-lattice, subcritical process admitting Malthusian parameter* $\alpha < 0$

$$m_t^a \sim e^{\alpha t} \int_0^a e^{-\alpha t}\{1 - L(t)\}dt \Big/ \int_0^\infty te^{-\alpha t} \, \mu(dt) \tag{4}$$

for $0 \leq a < \infty$, *as* $t \to \infty$. *For* $a = \infty$ *the relation still holds, providing*

$$\int_0^\infty te^{-\alpha t} L(dt) < \infty. \tag{5}$$

Comment. This result shows that asymptotically we expect the population count to change with the same speed as the exponential function. It is appropriate to mention here that the behaviour of the Galton–Watson processes have the same behaviour like GBP (see Slavchova–Bojkova et al. [4]), i.e. inside the GBP there is an embedded Galton–Watson process for the number of women belonging to the same generation.

1.2 GBP Model for Human Population

From human population point of view, the point process is the number of children a woman has in a particular interval. It is a random measure defined on the intervals (more exactly the Borel σ-algebra). For example, if a woman gives three births during her life then $\xi_x(\infty) = 3$. If two of them happened when she was between 20 and 30, then $\xi(20, 30) = 2$. It becomes obvious that $\xi(t)$ has only integer values but our expectation $\mu[a, b] = \mathbb{E}(\xi[a, b])$ can be non-integer. Another property of the human population is that children cannot give birth so $\xi(t) = 0$ and $\mu(t) = 0$ when $t < 12$. For a human population it is appropriate to consider a point processes $\xi(t)$, such that the corresponding $\mu(t)$ is smooth function. It is also appropriate to model the distribution $L(t) = 1 - S(t)$ of λ as smooth function. The function $S(t)$ is called survivability function of a live birth. We have $S(0) = 1$ and $S(\omega) = 0$, where ω is the oldest age in a life table. In this section we will follow the approach in [3] for modelling human population by GBP. In what follows the index x is skipped for simplicity. The model rest on the following assumptions:

i) The fertility interval for each woman is $[12, 50]$ and women cannot give birth outside it or if they are not alive. In terms of GBP

$$\mathbb{P}(\xi[a, b) = 0) = 1, \quad \text{when } [a, b) \cap [12, 50] = \varnothing,$$
$$\mathbb{P}(\xi[\lambda, \infty) = 0) = 1.$$

ii) A woman could have 0 or 1 daughter during a year and each birth is a live birth. This means the number of live births is equal to the number of women who gave birth.

iii) There is no migration.

Let ${}_bz_t$ is the branching process started from a woman aged b at time $t = 0$, ${}_b\xi$ is her point process, ${}_b\mu$ is the expectation of the point process and ${}_bS$ is her survivability function. Let $n_b = \mathbb{P}(\xi[b, b + 1) = 1 \mid \lambda \geq b)$ be the probability a woman gives birth at age b, if she survived to the beginning of this age interval. Let us shortly remind some results from [3], which we need for modelling purposes.

Theorem 3. *The expected population count at time t started from woman aged b at time zero is given by the following equation*

$$ {}_bm_t = {}_bS(t) + \int_0^t m_{t-u}\ {}_b\mu(b + du), \tag{6}$$

where ${}_bS(t)$ denotes the probability a woman of age b to survive to $b + t$.

To obtain approximation of the last equation an assumption for smoothness of $\mu(t)$ is made. Finding solution ${}_bm_t$ for equation (6) can be reduced to finding solution m_t for equation (3).

Lemma 1. *If m_t has a continuous second derivative, then a third order approximation of equation* (6) *is given by*

$$ {}_b m_t \approx {}_b S(t) + \sum_{k=1}^{n} m_{b+k-0.5} \cdot {}_b\mu(b+k-1, b+k). $$

The methodology for modelling human population is described in more details in [3].

2 Population Projections

When it comes to the reality the theoretical models often defer from it. In our model, for example, we suppose that the birth's and death's distributions do not change in time. However, in reality due to the advances in medicine, nowadays the risk of death especially for children is reducing. We can see a tendency of decreasement in children mortality. On the other hand, the urbanisation of societies is leading to decreasement of the number of children a woman wish give birth to. Moreover, we see that many factors are contributive for the population count and these factors change in time, so the distributions of interest must change too. However, this problem can be solved, as we will explain later.

It is possible to derive the birth and death probabilities for a particular year (from data just for that single year). These are called period probabilities as opposed to cohort probabilities of birth and death (see [5]). If we use the model to project just for one year ahead, then we can assume the distributions have not changed. If we have a particular expectation for the distributions of S and ξ for the following year, then we can use it to project for period $[1, 2]$ again assuming that distributions do not change in this interval. Then again using expectation for the third year we can make a one year projection to fourth year and so on. The implication of this idea in practice is that we gain:

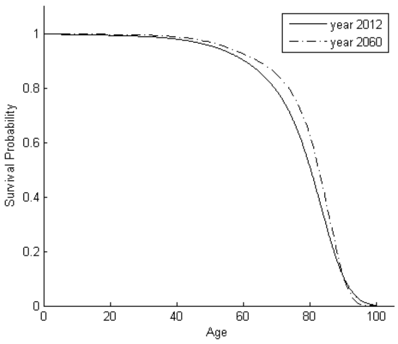
I) the ability to make a projection for the population count using our view of how S and ξ change through time;

II) the ability to distinguish population properties between years and see the effects of urbanisation, crisis or other social factors on these properties;

III) given a particular assumption for the demographic effect of some social policy of the government, we can see how it affects the future and is it more effective than some other policy.

We made a forecast of the birth and death distributions using EUROSTAT data (see [6]). The methodology can be shortly described in several steps:

1) deriving demographic rates (see [7]);
2) functional data analysis applying spline smoothing for derived probability distributions (see [8]). This step is done using R packages (see [9–11]);
3) principal component analysis of the distributions and forecasting using time series.



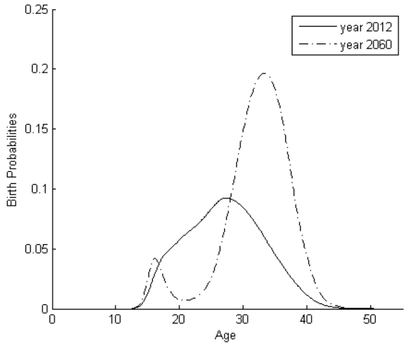


Fig. 1. The latest known survivability function (2012) against the expected future distribution (2062)

Fig. 2. The latest known birth density function (2012) against the expected future density function (2062)

In what follows we will consider two scenarios. The first will be that the current demographic state is stable and stays the same in the future. This scenario can be considered true for short time interval. The second scenario is that after the beginning of the third urbanization period in Bulgaria (after 1980) the demographic characteristics changed. Using data from 1980 to 2012 we fit a random walk with drift in order to capture the tendency and see what it predicts for the future. This scenario represents our expectation of what will happen given the social policy do not change.

In Figures 1 and 2 it is shown what the future will be if the tendency persists. The survival probabilities for females are increasing and the probabilities of birth seems to develop two peaks. Part of women prefers having children later in their life and part of them has children early. These results show a tendency that is observed in our society. Still the results for long periods ahead must be treated with caution, as they cannot take into account the human nature and culture. Forecasting the desire to have children is a difficult task that depends on many unpredictable factors like future financial crisis or war. It is observed for example that the number of children born during a war is increasing. This is an example that there are many natural processes that we know little about and cannot yet model good enough.

In Figure 3 we can see that the expected population counts, corresponding to the two scenarios do not differ for 20 years interval. That is due to the slow changes in distributions (again given there is no natural cataclysm or war). After that our expectation is that women start giving birth to more children and the expected population is better than the first scenario.

The Malthusian parameter estimated using the GBP model tells us that we expect the demographic condition to get better in time (Fig. 4).

The estimation of the GBP parameters is done in one year periods which is accurate if we are interested in the particular conditions during that year.

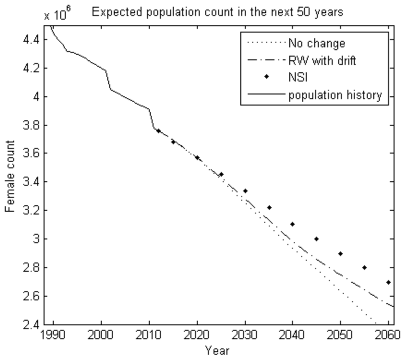


Fig. 3. Comparison of two scenarios of population development given no migration

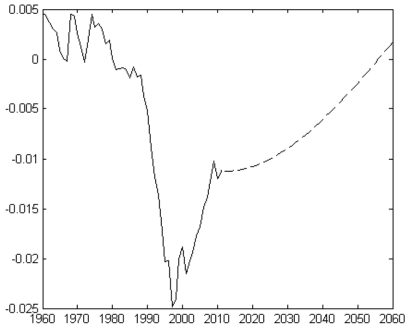


Fig. 4. Behaviour of the Malthusian parameter history (the solid line) and expected behaviour (the dashed line)

However, this may induce outliers in the birth's and death's distributions (caused by temporary social circumstances) which may bring some noise in the forecast. To avoid this imperfection, it is of interest to use data for several years to estimate the maximum likelihood parameters of the GBP and compare the forecasts in the future work. One useful approach to this problem can be the EM-algorithm for estimation of the offspring distribution in branching processes presented in [12]. Another interesting direction for future work is which of the social and economic conditions affect mostly the Malthusian parameter and the number of children a woman wants to have in her life. A possible approach is the probit model for one continuous and one ordinary variable which may be estimated with the EM-algorithm developed in [13].

3 Conclusions

The results for the expected future population according to different scenarios seem close to each other. Even though the forecasts of the birth and death probabilities are different in each scenario, the age structure of the population is similar in short term forecasts. We can see that the tendency in the past 30 years is resulting in an optimistic forecast that the Malthusian parameter will increase and the GBP will slowly change from subcritical to critical. The projection of the Bulgarian National Statistical Institute (NSI) based on assumption of convergence in the EU shows slightly more optimistic expectation for the population count than the projection based on the tendency in the past 30 years. However, in this debt crisis the social differences between the countries are increasing, especially between the core and the periphery. Thus the convergence assumption of NSI and EUROSTAT is put to test.

Acknowledgements

This work is partially supported by funds of the National Project for Supporting Doctoral Students and Young Scientists in Mathematics and Computer Science at Sofia University "St. Kliment Ohridski" and co-financed by the European Social Fund through the Human Resource Development Operational Programme under contract BG051PO001-3.3.06-0052 (2012/2014).

References

1. Jagers, P.: Branching Processes with Biological Applications. John Wiley and Sons Ltd. (1975)
2. National Statistical Institute. `http://www.nsi.bg`
3. Trayanov, P.I.: Crump–Mode–Jagers Branching Process: Modelling and Application for Human Population. Pliska Stud. Math. Bulgar. **22** (2013) 207–224
4. Slavtchova–Bojkova, M., Yanev, N.M.: Branching Stochastic Processes. St. Kliment Ohridski University Press, Sofia (2007)
5. Keyfitz, N., Caswell, H.: Applied Mathematical Demography. Springer (2005)
6. Eurostat Database. `http://epp.eurostat.ec.europa.eu/portal/page/portal/statistics/search_database`
7. Human Mortality Database. University of California, Berkley, Max Planck Institute for Demographic Research. `http://www.mortality.org`
8. Ramsay, J.O., Silverman, B.W.: Functional Data Analysis. Springer (2005)
9. R Development Core Team: A Language and Environment for Statistical Computing. (2011). `http://www.R-project.org/`
10. Hastie, T.: Generalized Additive Models. (2011) R package version 1.09-1. `http://CRAN.R-project.org/package=gam`
11. Hyndman, R.J., Booth, H., Tickle, L., Maindonald, J.: Forecasting Mortality, Fertility, Migration and Population Data. (2011) R package version 1.09-1. `http://CRAN.R-project.org/package=demography`
12. Daskalova, N.: Using Inside-Outside Algorithm for Estimation of the Offspring Distribution in Multitype Branching Processes. Serdica Journal of Computing **4**(4) (2010) 463–474
13. Grigorova, D., Gueorguieva, R.: Implementation of the EM algorithm for maximum likelihood estimation of a random effects model for one longitudinal ordinal outcome. Pliska Stud. Math. Bulgar. **22** (2013) 41–56

Using Classification Methods to Predict Market Demand for Products with Short Sales History

Nina Daskalova and Dragoslav Dragiev

Faculty of Mathematics and Informatics, Sofia University “St. Kliment Ohridski”,
5, James Bourchier Blvd., 1164 Sofia, Bulgaria
ninad@fmi.uni-sofia.bg

Abstract. Traditional approach for forecasting the demand in the market requires huge amount of data for long period of time. But in many areas in the market there is not enough data to make a good forecast. For example, different type of computers, cell phones and accessories are sold for short period of time and after that they are replaced with new models. Retailers need a good forecast of the demand for these new products. One approach is to use products with similar behaviour in the market to obtain segmentation that could be used for establishing demand profiles. Bayesian methods are also appropriate in such situations. In this paper, classification and Bayesian methods are proposed to find similar products and group them into clusters, which could be used to make prediction of the market demand for the product profiles.

Keywords: demand forecasting, classification methods, cluster analysis, Bayesian classification

1 Introduction

It is very important in retail industry to obtain a good forecast of the demand of certain merchandise. Regarding the demand as random, different statistical techniques proved themselves suitable to treat this problem. Methods from the domains of Bayesian analysis and statistical learning are commonly used in this field. For a comprehensive review of these methods see Gelman et al. [3], Hastie et al. [4], Mitchell [6] and West and Harrison [12].

Some products such as sport apparel, technical goods or fashion accessories are characterized by a high degree of demand uncertainty. New designs enter the market every season in this category. Because of the long lead times and relatively short selling seasons of such products retailers are forced to order well in advance of the sales season. Sometimes substantial amount of uncertainty about the demand process can be resolved using the early sales information. Dynamic pricing of a given item is successfully modelled when the demand is probabilistic and price sensitive. Most of inventory models that use a Bayesian approach assume that the demand in one period is random with a known distribution but with an unknown parameter. This unknown parameter has a prior probability distribution which reflects the initial estimates of the decision maker. Observed sales are then used to find the posterior distribution of the unknown parameter. This is called demand learning. Some of the basic research in this field could be seen in Scarf [7], Azoury [1], Iglehart [5] and Fisher and Raman [2]. A dynamic pricing model that incorporates demand learning is developed in Sen et al. [8].

To apply demand learning models, some initial information about the early stage of sales is needed. The estimation of the demand parameters usually does not perform well in this stage. It is useful for the decision maker to have some estimated forecast for the early sales. Often this is done using sales history of similar merchandise ordered in previous years, which retailers usually do assuming their personal knowledge and experience. An automated approach to this problem is aggregation of sales by items families or by clustering procedures to obtain complete historical data of several years. An application to textile industry using different soft computing techniques like neural networks and decision trees is considered in the works of Thomassey et al. [9–11]. In this paper we propose a simpler statistical approach, based on cluster methods and Bayesian classification, that is not computationally expensive and could be used in practice in different fields. We have tested it on a dataset for sports apparel.

The paper is organized as follows. Section 1 introduces the data available and discusses some special features. In Section 2 k-means and hierarchical clustering algorithms are reviewed. These are used to group our training sample in clusters. Later, these clusters will be used to predict future market demand for new products in Sect. 3. The classification method which proved to be most suitable in this case – the Naive Bayes Classifier (NBS) is introduced in Sect. 4. The last two sections present the results and the conclusions of our analysis.

2 Data Description

Available data contains sales for 61 different products. Every product (x_i, $i = 1, \ldots, 61$) is observed for 69 consecutive weeks. A transformation is performed that moves the zeros from the beginning to the end of the time series for every object, so it starts with a nonzero sale entry. Also, sales are recomputed in percentage of the whole sale for the period. This ensures reasonable comparison of the sales as vectors for different products. We regard these vectors as products' *sales profiles*. The profiles of some of the products are given in Fig. 1.

The use of these profiles is very natural, for example it distinguishes the fast sellers from slower ones, and retail professionals recognize several types of profiles of the sales life-cycle. For example the fifth profile in the first line and the second in the second line are typical for fast-selling products and the last one in the first line is an obvious slow-seller. In the third line three profiles of products with double peak in sales are shown. This usually is due to some price discount. The first task in our analysis is to divide products' profiles into groups using cluster methods. The clusters may or may not overlap the pre-defined groups of "fast-sellers", "slow-sellers" and "double peaks". The existence of some overlapping will show that the clustering is reasonable. Then, a new product will be classified into one of the existing clusters using NBC. The mean of the sales

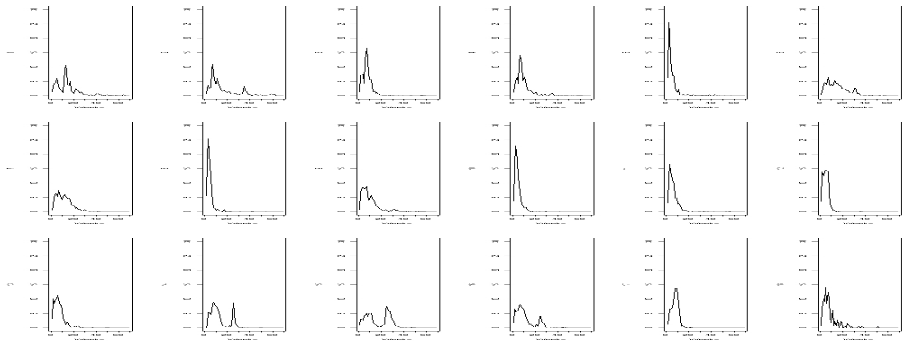


Fig. 1. Sales profiles of some of the products in the dataset

of the products in this cluster could be used as a prediction for the sales of that product for the coming week. The procedure has to be repeated for every week of the sale-cycle.

3 Cluster Analysis

Cluster analysis, also called data segmentation, is used to group a collection of objects into subsets or *clusters*, such that those within each cluster are more closely related to one another than objects assigned to different clusters. The goal sometimes is to arrange the clusters into a natural hierarchy. An important notion is to determine the degree of similarity between the individual objects being clustered. A clustering method attempts to group the objects based on the definition of similarity supplied to it.

3.1 K-means Algorithm

The K-means algorithm is one of the most popular iteratively descent clustering methods. It is intended for situations in which all variables are of quantitative type, and squared euclidian distance $d(x_i, x_{i'}) = \Sigma_{j=1}^{p}(x_{ij} - x_{i'j})^2 = \|x_i - x_{i'}\|^2$ is chosen as the dissimilarity measure. The within point scatter can be written as:

$$W(C) = \frac{1}{2}\Sigma_{k=1}^{K}\Sigma_{C(i)=k}\Sigma_{C(i')=k}\|x_i - x_{i'}\|^2 = \sum_{k=1}^{K} N_k \sum_{C(i)=k} \|x_i - \bar{x}_k\|^2,$$

where $\bar{x}_k = (\bar{x}_{1k}, \bar{x}_{2k}, \ldots, \bar{x}_{pk})$ is the mean vector associated with k-th cluster and $N_k = \sum_{i=1}^{N} I(C(i) = k)$. To define the number of clusters, the within cluster dissimilarity W_K is given as a function of the number of clusters K in Fig. 2.

This shows that the best choice for the number of clusters is 3 or 4. The results in Fig. 3 are for 4 clusters. Average values are computed and are added with black line.

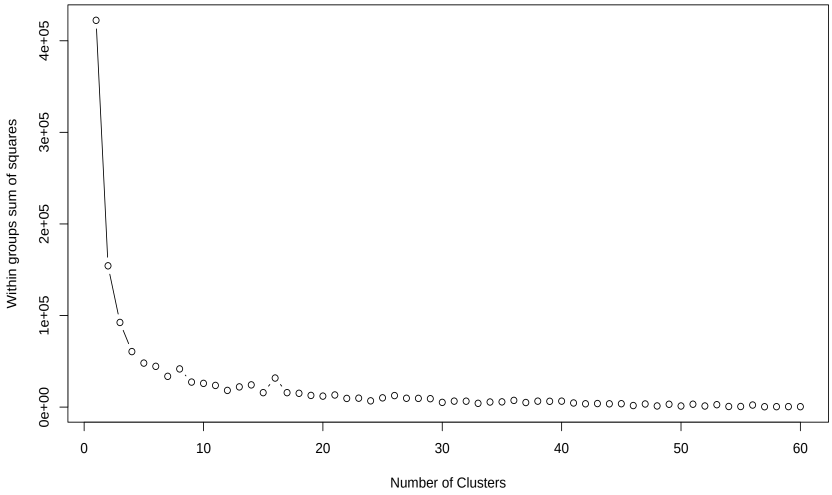


Fig. 2. Scree plot of clusters

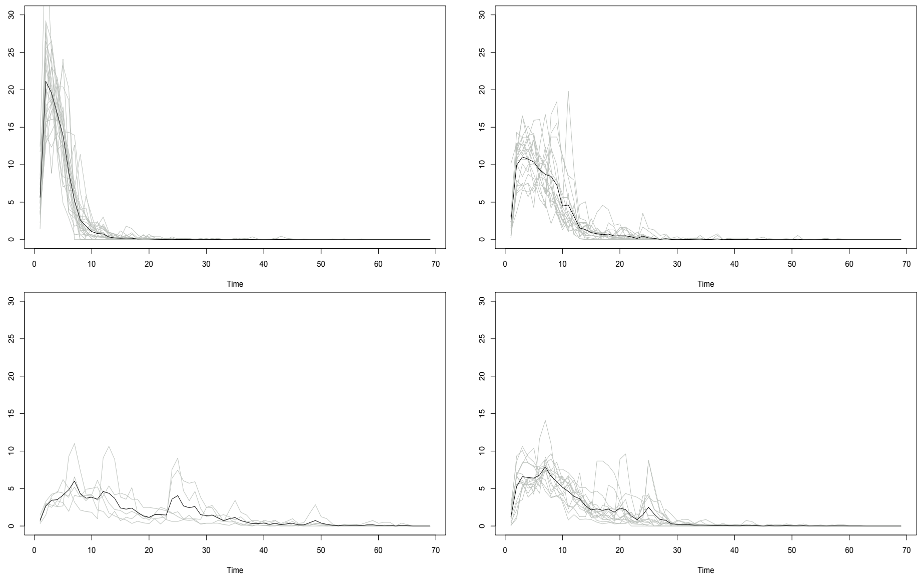


Fig. 3. The four clusters with their mean (in black)

3.2 Hierarchical Cluster Analysis

The result of applying K-means clustering algorithm depends on the choice for the number of clusters to be searched and a starting configuration assignment. In contrast, hierarchical clustering methods do not require such specifications. Instead, they require the user to specify a measure of dissimilarity between (dis-

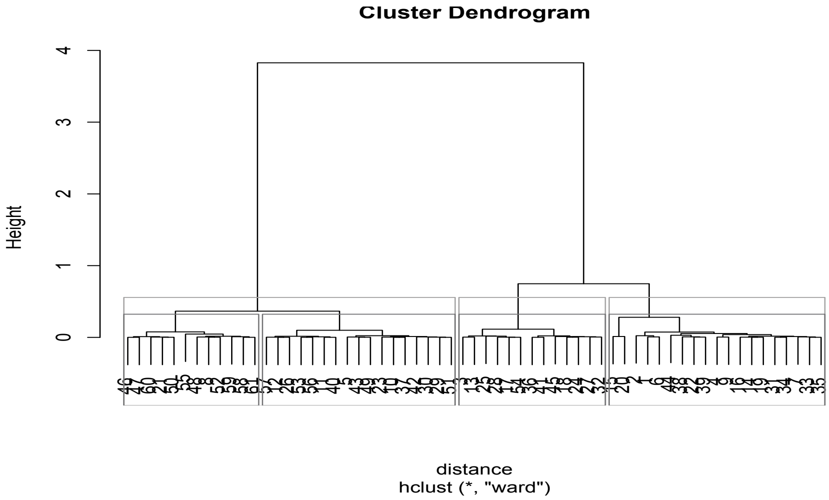


Fig. 4. Hierarchical clustering

joint) groups of observations, based on the pairwise dissimilarities among the observations in the two groups. A measure of pairwise dissimilarity among the observations is chosen:

$$1 - \rho(x_i, x_{i'}) = 1 - \frac{\sum_j (x_{ij} - \bar{x}_i)(x_{i'j} - \bar{x}_{i'})}{\sqrt{\sum_j (x_{ij} - \bar{x}_i)^2 \sum_j (x_{i'j} - \bar{x}_{i'})^2}}.$$

As the name suggests, the algorithm produces hierarchical representations in which the clusters at each level of the hierarchy are created by merging clusters at the next lower level. It is up to the user to decide which level (if any) actually represents a "natural" clustering. Recursively binary splitting/agglomeration can be represented by a rooted binary tree, called *dendrogram*. For the given data, Figure 4 shows the dendrogram resulting from hierarchical clustering with Ward clustering procedure (clusters are generated that minimize the squared Euclidean distance to the center mean) and squared Euclidean distance.

In this case we have four clusters with 19, 13, 17, 12 observations. Graphically our clusters are given in Fig. 5. Average values are computed and are added with black line.

The four clusters suggested by the k-means algorithm look closer to the pre-defined groups of profiles, thus having better practical explanation. For this reason k-means has been chosen in our application.

4 Naive Bayes Classifier

The Naive Bayes Algorithm is a classification algorithm based on Bayes' rule, that assumes that attributes $X_1, X_2, \ldots, X_n$ are all conditionally independent

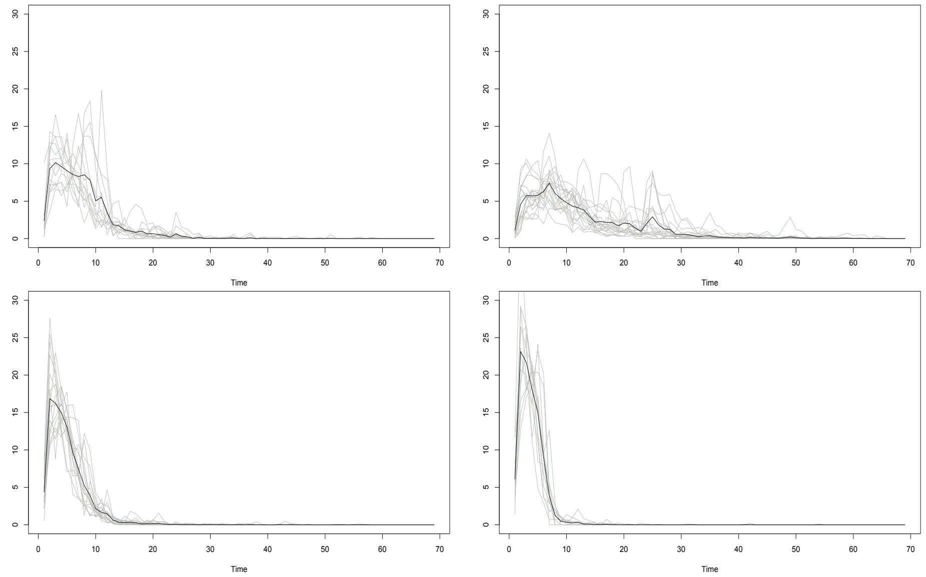


Fig. 5. The four clusters with their mean (in black)

of one another, given Y: $\Pr(X_1, X_2, \dots, X_n|Y) = \prod_{i=1}^{n} \Pr(X_i|Y)$. Our goal is to train a classifier that will output the probability distribution over possible values of Y, for each new instance of X that we ask it to classify. The expression for the probability that Y will take on its k-th possible value, according Bayes' rule, is as follows:

$$\Pr(Y = y_k|X_1, \dots, X_n) = \frac{\Pr(Y = y_k) \prod_{i=1}^{n} \Pr(X_i|Y = y_k)}{\sum_j \Pr(Y = y_j) \prod_{i=1}^{n} \Pr(X_i|Y = y_j)}.$$

If we are interested only in the most probable value of Y, then we have the Naive Bayes Classification Rule:

$$Y \leftarrow \text{argmax}_{y_k} \frac{\Pr(Y = y_k) \prod_{i=1}^{n} \Pr(X_i|Y = y_k)}{\sum_j \Pr(Y = y_j) \prod_{i=1}^{n} \Pr(X_i|Y = y_j)},$$

which simplifies to the following

$$Y \leftarrow \text{argmax}_{y_k} \Pr(Y = y_k) \prod_{i=1}^{n} \Pr(X_i|Y = y_k)$$

because the denominator does not depend on y_k.

This will be used to classify any new product to one of the pre-defined clusters of profiles. This step will be repeated for every week of the sales cycle to obtain the "predicted" product's profile. The NBC have been chosen as a method for classification, because it only requires a small amount of training data to estimate the parameters. The scarcity in the data justified the acceptance of the naive and oversimplified assumptions of this method. Compared to other simple classification methods (such as linear discriminant analysis) it showed better behaviour for our data.

5 Results

Making a prediction for a new product without any sales information requires using different characteristics of this product and other products from the training sample (or an expert opinion) to find similar products and classifying it in the correct cluster. Usually the new products are in fact successors of known products from previous seasons and it is easy to define the prospective sales profile. Knowing product sales for i weeks, starting from the first week, we use Naive Bayes Classifier (NBC) to classify the product to a particular cluster and use the mean in this cluster as a prediction for the future demand. Receiving information for week $i+1$ $(i+2,\dots)$, we use again NBC to check if this product stays in this cluster or if it moves to another and again use mean in the corresponding cluster to make a prediction for future demand.

Cross-validation results using NBC week by week are shown in Fig. 6. The graphics compare estimated sales (dashed line) for a given product to the real sales (solid line). The training sample consists of 60 products. Computations are implemented in R.

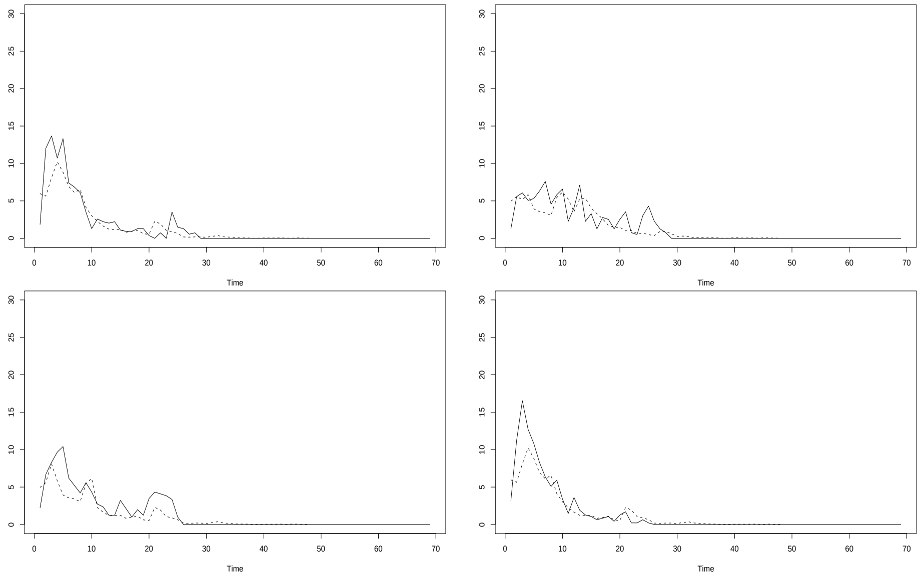


Fig. 6. The resulting predicted profile vs. real one for some of the products in the test set

6 Conclusion

The method proposed showed to be easy to implement and effective enough to use in case of products with short or no sales history. In the test set available it determined very reasonable profile clusters, and performed well in classifying new products after the first few weeks. Further incorporating with other forecasting methods is considered to improve performance, especially in the beginning of the sales cycle.

Acknowledgements

This work was supported by the European Social Fund through the Human Resource Development Operational Programme under contract BG051PO001-3.3.06-0052 (2012/2014).

References

1. Azoury, K.S.: Bayes solution to dynamic inventory models under unknown demand distributions. Management Science **31** (1985) 1150–1161
2. Fisher, M., Raman, A.: Reducing the cost of demand uncertainty through accurate response to early sales. Operations Research **44** (1996) 87–99
3. Gelman, A., Carlin, J.B., Stern, H.J., Rubin, D.B.: Bayesian Data Analysis, 2nd Edition. Chapman & Hall/CRC Press (2004)
4. Hastie, T., Tibshirani, R., Freedman, J.: The elements of Statistical Learning, 2nd Edition. Springer, New York (2009)
5. Iglehart, D.L.: The dynamic inventory problem with unknown demand distribution. Management Science **10** (1964) 429–440
6. Mitchell, T.M.: Machine Learning. McGraw Hill, New York (2010)
7. Scarf, H.: Bayes solutions to the statistical inventory problem. Annals of Mathematical Statistics **30** (1959) 490–508
8. Sen, A., Zhang, A.X.: Style goods pricing with demand learning. European Journal of Operational Research **196**(3) (2009), 1058–1075
9. Thomassey, S., Happiette, M., Dewaele, N., Castelain, J.M.: A Short and Mean Term Forecasting System Adapted to Textile Items' Sales. Journal of the Textile Institute **93**(3) (2002) 95–104
10. Thomassey, S, Fiordaliso, A.: A hybrid sales forecasting system based on clustering and decision trees. Decision Support Systems **42** (2006) 408–421
11. Thomassey, S., Happiette, M.: A neural clustering and classification system for sales forecasting of new apparel items. Applied Soft Computing **7** (2007) 1177–1187
12. West M., Harrison J.: Bayesian Forecasting and Dynamic Models, 2nd Edition. Springer, New York (1999)
13. R Development Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. (2010) http://www.R-project.org

EM Algorithms for MLE of Correlated Probit Models

Denitsa Grigorova[1] and Ralitza Gueorguieva[2]

[1] Faculty of Mathematics and Informatics, Sofia University "St. Kliment Ohridski",
5, James Bourchier Blvd., 1164 Sofia, Bulgaria
dgrigorova@fmi.uni-sofia.bg

[2] Yale University, School of Public Health, Department of Biostatistics,
New Haven, Connecticut, USA
ralitza.gueorguieva@yale.edu

Abstract. The probit model is frequently used for modeling of univariate ordered data. Its main feature is the assumption of a latent normal variable which determines the level of the observed ordinal response via thresholds. The correlated probit model is an extension of the probit model to multiple outcomes. It allows to seamlessly incorporate different correlation structures of the latent variables related to the observed ordinal outcomes. The estimation of the correlated probit model parameters based on direct maximization of the limited information maximum likelihood is a numerically intensive procedure. In this manuscript we review two EM algorithms for obtaining Maximum Likelihood Estimates (MLE) for correlated probit models. The first correlated probit model we consider is for one longitudinal ordinal outcome. The other correlated probit model is for multiple ordinal outcomes. The algorithms are implemented in the free software environment for statistical computing and graphics **R** [1]. We present a simulation study to examine the reliability of the EM algorithm for MLE of correlated probit model for multiple ordinal outcomes.

Keywords: correlated probit model, EM algorithm, random effects

1 Introduction

The probit model was first introduced by Bliss [2, 3]. Its main feature is the assumption of a latent variable which determines the level of the observed ordinal response via thresholds. The usefulness of the model is not affected when the existence of the latent variable does not seem natural. Extensions of the classical probit model are correlated probit models [4]. They are widely used for modeling of multiple categorical variables or clustered/longitudinal ordinal outcomes because these models have two main advantages. They are easy to interpret and they allow rich correlation structure of the latent variables via random effects and/or correlated errors. This allows to take into account the natural dependence of the measurements on the same subject.

The correlated probit model does not have closed form expression for the likelihood function. Approximations need to be used in order to obtain estimates of the unknown parameters. There are several methods of statistical inference based on numerical, stochastic or analytical approximations. Most popular are extensions of numerical approximations such as Gauss-Hermite quadrature [5, pp. 306–307] or adaptive Gaussian Quadrature [6]. Another approach is based on analytical approximations (Breslow and Clayton [7], Wolfinger and O'Connell [8]) but it has been shown to produce bias in the parameter estimates especially for binary data or ordinal data with few categories. A third approach is the Expectation-Maximization (EM) algorithm [9].

The EM algorithm was given its name by Dempster, Laird and Rubin [9]. They develop a general framework of the EM algorithm which before them was proposed by other authors in special circumstances. The EM algorithm is an iterative procedure for finding of Maximum Likelihood Estimates (MLE) in models which depend on unobserved data. The algorithm consists of two steps: E-step (Expectation-step) and M-step (Maximization-step). An extension of the EM algorithm is the Expectation/Conditional Maximization (ECM) algorithm [10] which replaces the complicated in some cases M-step of the EM algorithm with several computationally simpler CM-steps.

The first to apply the EM algorithm to probit models is Ruud [11]. Kawakatsu and Largey [12] extend his approach to a model for one ordinal and multiple normal outcomes. We combine their results and the approach of Chan and Kuk [13] to development of ECM algorithms for two correlated probit models. The first one is for one longitudinal ordinal response. The second correlated probit model is for multiple ordinal outcomes.

The paper is organized as follows. Section 2 defines the two correlated probit models and outlines the idea of the EM algorithms for estimation in the correlated probit model for multiple ordinal responses. Section 3 describes a simulation study that was performed in order to examine the reliability of the algorithm. Section 4 contains concluding remarks and discussion about possible extensions of the algorithms.

2 Models

2.1 Model for One Longitudinal Ordinal Outcome

Let y^*_{ij} denote the observed ordinal variable with m levels on the i-th subject at time j. We assume that there is a latent normal variable y_{ij} that generated the observed variable. We consider the following random effects model:

$$y_{ij} = \boldsymbol{x'_{ij}}\boldsymbol{\beta} + \boldsymbol{z'_{ij}}\boldsymbol{b_i} + \epsilon_{ij}. \tag{1}$$

The rule that relates the latent variable to the observed ordinal variable is:

$$y^*_{ij} = \begin{cases} 1, & y_{ij} \le \alpha_1, \\ j, & \alpha_{j-1} < y_{ij} \le \alpha_j,\ j = 2, \dots, m-1, \\ m, & y_{ij} > \alpha_{m-1} \end{cases} \tag{2}$$

for some unknown thresholds $\alpha_1, \dots, \alpha_{m-1}$.

The vector of random effects is assumed to be normally distributed, q-dimensional and is denoted by $\boldsymbol{b_i} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$. The covariance matrix $\boldsymbol{\Sigma}$ is a quadratic $q \times q$ positive semi-definite matrix. The error term is normally distributed $\epsilon_{ij} \sim N(0, \sigma^2)$ and is independent of the random effects.

The regression parameters of the fixed effects in model (1) are denoted by the p-dimensional vector $\boldsymbol{\beta}$. The vector of predictors for the fixed effects is $\boldsymbol{x_{ij}}$ and the vector of predictors for the random effects is $\boldsymbol{z_{ij}}$.

From the observed data we can not estimate all of the unknown parameters. One possibility for identifiability restrictions is the following: the first threshold α_1 is set to zero and the variance of the normal error term σ^2 is set to 1. Other restrictions and re-parametrization are possible.

2.2 Model for Multiple Ordinal Outcomes

We observe p ordinal outcomes on the same subject i with respectively m_1, m_2, $\dots$, m_p levels denoted by $\boldsymbol{y_i^*} = (y_{i1}^*, y_{i2}^*, \dots, y_{ip}^*)'$. We assume that latent normal variables y_{ij}, $j = 1, 2, \dots, p$ generate the observed variables. We consider the following correlated probit model for the latent variables:

$$y_{ij} = \boldsymbol{x'_{ij}\beta_j} + z_{ij}b_{ij} + \epsilon_{ij}, \quad j = 1, 2, \dots, p \tag{3}$$

where we observe $y_{ij}^* = l_j$, $l_j = 2, \dots, m_j - 1$ if $\alpha_{j,l_j-1} < y_{ij} \le \alpha_{j,l_j}$, $y_{ij}^* = 1$ if $y_{ij} \le \alpha_{j,1}$ and $y_{ij}^* = m_j$ for $y_{ij} > \alpha_{j,m_j-1}$ for some thresholds $\alpha_{j,1}, \dots, \alpha_{j,m_j-1}$, $j = 1, 2, \dots, p$.

We assume a normal distribution of the p-dimensional vector of random intercepts $\boldsymbol{b_i} = (b_{i1}, \dots, b_{ip})' \sim N(\mathbf{0}, \boldsymbol{\Sigma})$. The covariance matrix $\boldsymbol{\Sigma}$ is a quadratic $p \times p$ positive semi-definite matrix. The error terms are independent normally distributed $\epsilon_{ij} \sim N(0, \sigma^2)$. We also assume that random effects and error terms are independent of each other.

Regression parameters for the fixed effects in model (3) are denoted by q_j-dimensional vectors $\boldsymbol{\beta_j}$, $j = 1, \dots, p$. The vectors of the predictors for the fixed effects are $\boldsymbol{x_{ij}}$, $j = 1, \dots, p$ and the predictors for the random intercepts are z_{ij}, $j = 1, \dots, p$. From the observed data it is not possible to uniquely estimate all of the unknown parameters. We pose the following identifiability restrictions: the first thresholds $\alpha_{j,1}$, $j = 1, \dots, p$ are zero and the variance of the normal error terms σ^2 is 1. Some other restrictions and re-parametrization are possible. Next section provides description of the developed ECM algorithm for the model.

2.3 EM Algorithms for MLE of Both Models

The idea of both algorithms is the same. The details below are description of the EM algorithm for model (3). Details of the EM algorithm for the correlated probit models for one longitudinal ordinal response can be found in Grigorova and Gueorguieva [14]. The first step is to re-parametrize the thresholds so that they can be explicitly included in the complete data log-likelihood. For this purpose we use the approach of Kawakatsu and Largey [12] who extend Ruud's work [11]. According to their method we define the differences between consecutive thresholds with $\delta_{j,i} = \alpha_{j,i} - \alpha_{j,i-1}$, $i = 2, \ldots, m_j - 1$, $j = 1, 2, \ldots, p$ (we define additionally $\delta_{j,1} = \delta_{j,m_j} = 1$). It follows that $\alpha_{j,i} = \sum_{k=2}^{i} \delta_{j,k}$, $j = 1, 2, \ldots, p$, $i = 2, \ldots, m_j - 1$. Then we consider new variables that are linear transformations of the latent variables. The new variables are denoted by $y_{ij_{\text{new}}} = (y_{ij} - \alpha_{j,y^*_{ij}-1})/\delta_{j,y^*_{ij}}$, $j = 1, 2, \ldots, p$, where $\alpha_{j,0} = 0$, $j = 1, 2, \ldots, p$ and $\boldsymbol{y_{i_{\text{new}}}} = (y_{i1_{\text{new}}}, y_{i2_{\text{new}}}, \ldots, y_{ip_{\text{new}}})'$. For example if $y^*_{ij} = u$ then $y_{ij_{\text{new}}} = (y_{ij} - \alpha_{j,u-1})/\delta_{j,u}$. Since the new variables are linear transformations of the latent variables then they have normal distributions. But conditional on the observed variables, the transformed variables have truncated multivariate normal distribution.

If we observe the first level of y^*_{ij}, the new variable $y_{ij_{\text{new}}}$ is truncated at $(-\infty, 0]$. If y^*_{ij} is between the first and the last level, the new variable is truncated at $(0, 1]$. If we observe the last level of y^*_{ij}, the new variable is truncated at $(0, \infty)$.

We use the approach of Chan and Kuk [13] in order to find closed form expressions for the unknown parameters $\boldsymbol{\Gamma} = (\boldsymbol{\beta'_1}, \boldsymbol{\beta'_2}, \ldots, \boldsymbol{\beta'_p}, \boldsymbol{\Sigma}, \boldsymbol{\delta'_1}, \boldsymbol{\delta'_2}, \ldots, \boldsymbol{\delta'_p})$, where $\boldsymbol{\delta_j} = (\delta_{j,2}, \ldots, \delta_{j,m_j-1})$, $j = 1, \ldots, p$.

Complete Data Log-likelihood. The complete data log-likelihood has the following form:

$$\ln L = \ln f(\boldsymbol{b}, \boldsymbol{y_{\text{new}}}) = \sum_{i=1}^{n} \ln f(\boldsymbol{b_i}) f(\boldsymbol{y_{i_{\text{new}}}}|\boldsymbol{b_i}) = \sum_{i=1}^{n} \ln f(\boldsymbol{b_i}) \prod_{j=1}^{p} f(y_{ij_{\text{new}}}|\boldsymbol{b_i}).$$

Apart from the constants the log-likelihood is as follows:

$$\begin{aligned}\ln L = &- 0.5 \sum_{i=1}^{n} \ln |\boldsymbol{\Sigma}| - 0.5 \sum_{i=1}^{n} \boldsymbol{b'_i} \boldsymbol{\Sigma^{-1}} \boldsymbol{b_i} \\ &+ \sum_{i=1}^{n} \ln \delta_{1,y^*_{i1}} - 0.5 \sum_{i=1}^{n} [\delta_{1,y^*_{i1}} y_{i1_{\text{new}}} - (\boldsymbol{x'_{i1}} \boldsymbol{\beta_1} + z_{i1} b_{i1} - \alpha_{1,y^*_{i1}-1})]^2 + \cdots \\ &+ \sum_{i=1}^{n} \ln \delta_{p,y^*_{ip}} - 0.5 \sum_{i=1}^{n} [\delta_{p,y^*_{ip}} y_{ip_{\text{new}}} - (\boldsymbol{x'_{ip}} \boldsymbol{\beta_p} + z_{ip} b_{ip} - \alpha_{p,y^*_{ip}-1})]^2.\end{aligned}$$

Closed Form Expressions for the Estimators. We obtain closed form expressions for the estimators of the unknown parameters by setting the first derivatives of the complete data log-likelihood to zero.

The estimator for the covariance matrix $\boldsymbol{\Sigma}$ of the random effects is:

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n} \sum_{i=0}^{n} \boldsymbol{b_i} \boldsymbol{b'_i}.$$

Regression parameters for the fixed effects $\boldsymbol{\beta_j}$, $j = 1, 2, \ldots, p$ satisfy the following system of equations:

$$\sum_{i=1}^{n} \boldsymbol{x_{ij}}\boldsymbol{x'_{ij}}\boldsymbol{\beta_j} = \sum_{i=1}^{n}[\delta_{j,y^*_{ij}} y_{ij_{\text{new}}} - z'_{ij} b_{ij} + \alpha_{j,y^*_{ij}-1}]\boldsymbol{x_{ij}}.$$

Therefore the regression parameters $\boldsymbol{\beta_j}$ are a least square solution of the regression of $\tilde{y}_{ij}$ on $\boldsymbol{x_{ij}}$, where $\tilde{y}_{ij} = \delta_{j,y^*_{ij}} y_{ij_{\text{new}}} - z'_{ij} b_{ij} + \alpha_{j,y^*_{ij}-1}$, $j = 1, 2, \ldots, p$.

The equations for $\delta_{j,k}$, $k = 2, \ldots, m_j - 1$, $j = 1, 2, \ldots, p$ are quadratic equations of the form: $a_j \delta^2_{j,k} + b_j \delta_{j,k} + c_j = 0$, which always have real roots and the bigger root is always positive. The constants a_j, b_j, c_j are as follows:

$$\begin{aligned} a_j &= \sum_i \sum_{y^*_{ij}=k} (y^2_{ij_{\text{new}}}) + n_{j,k+1} + \cdots + n_{j,m}, \\ b_j &= -\sum_i \sum_{y^*_{ij}=k} y_{ij_{\text{new}}}(\boldsymbol{x'_{ij}}\boldsymbol{\beta_j} + z_{ij} b_{ij} - \alpha_{j,k-1}) \\ &\quad + \sum_i \sum_{y^*_{ij}>k} (\delta_{j,y^*_{ij}} y_{ij_{\text{new}}} - \boldsymbol{x'_{ij}}\boldsymbol{\beta_j} - z_{ij} b_{ij} + \alpha_{j,-k}), \\ c_j &= -n_{j,k}, \end{aligned}$$

where $n_{j,k}$ is the number of the observations of the categorical variable y_j at k-th level and $\alpha_{j,-k} = \delta_{j,2} + \cdots + \delta_{j,k-1} + \delta_{j,k+1} + \cdots + \delta_{j,y^*_{ij}-1}$.

In order to update the parameter estimates we need to express the conditional expectations given observed data in the closed form expressions for the estimators. It can be shown that all conditional expectations depend only on the first two moments of the truncated multivariate normal distribution.

$(k+1)$-st Iteration of the EM Algorithm. We use an extension of the EM algorithm called Expectation/Conditional Maximization algorithm [10]. We will write down the estimates of the unknown parameters at the $(k+1)$-st step of the EM algorithm, where $\boldsymbol{\Gamma^k}$ is the (k)-th estimate of the unknown parameters $\boldsymbol{\Gamma}$:

- The $(k+1)$-st estimate of the regression parameters $\boldsymbol{\beta^{k+1}_j}$, $j = 1, 2, \ldots, p$ is a least square solution of the regression of $E(\tilde{y}_{ij}|\boldsymbol{y^*_i};\boldsymbol{\Gamma^k})$ on $\boldsymbol{x_{ij}}$.
- The $(k+1)$-st estimate of $\delta_{j,u}$, $u = 2, \ldots, m_j - 1$, $j = 1, 2, \ldots, p$ is

$$\delta^{k+1}_{j,u} = \frac{-E[b_j|\boldsymbol{y^*};\boldsymbol{\Gamma^k}] + \sqrt{(E[b_j|\boldsymbol{y^*};\boldsymbol{\Gamma^k}]^2 - 4E[a_j|\boldsymbol{y^*};\boldsymbol{\Gamma^k}]E[c_j|\boldsymbol{y^*};\boldsymbol{\Gamma^k}])}}{2E[a_j|\boldsymbol{y^*};\boldsymbol{\Gamma^k}]}$$

  and in the expression for a_j, b_j, c_j we use the updated estimates $\boldsymbol{\beta^{k+1}_j}$, $\delta^{k+1}_{j,i}$, $i = 2, \ldots, u - 1$.
- The new estimate of the covariance matrix of the random effects is $\hat{\boldsymbol{\Sigma}}^{k+1} = \frac{1}{n}\sum_{i=0}^{n} E(\boldsymbol{b_i b'_i}|\boldsymbol{y^*_i};\boldsymbol{\Gamma^k})$. For the calculations of the expectations we use the updated in the previous steps estimates $\boldsymbol{\beta^{k+1}_j}$, $\delta^{k+1}_{j,i}$, $i = 2, \ldots, m_j - 1$, $j = 1, \ldots, p$.

3 Simulation Study

We simulated values from the following correlated probit model for two ordinal outcomes each with three levels:

$$\begin{aligned} y_{i1} &= \beta_{10} + \beta_{11}x_{i1} + b_{i1} + \epsilon_{i1}, \\ y_{i2} &= \beta_{20} + \beta_{21}x_{i2} + b_{i2} + \epsilon_{i2}, \end{aligned}$$

where $\beta_{10} = -0.5$, $\beta_{11} = 1$, $\beta_{20} = 1$, $\beta_{21} = -0.5$, $\mathrm{Var}(\epsilon_{ij}) = 1$, $j = 1, 2$, with thresholds $\alpha_{1,1} = \alpha_{2,1} = 0$, $\alpha_{1,2} = 1.2$, $\alpha_{2,1} = 0.7$ and

$$\mathrm{Var}\begin{pmatrix} b_{i1} \\ b_{i2} \end{pmatrix} = \begin{pmatrix} 1 & -0.8 \\ -0.8 & 1 \end{pmatrix}.$$

We simulated 100 samples with a sample size of $n = 500$ subjects. We used a bootstrap method for standard errors approximation [15, pp. 130–131]. The steps for model (1) are described in Grigorova and Gueorguieva [14] and for model (3) the procedure is analogous. For each approximation of the standard errors we used 50 bootstrap samples which is within the recommended range of 50 to 100 bootstrap replications (Efron and Tibshirani [16]). The results are presented in Table 1.

Table 1. Table of estimates and standard errors in the simulation study

Parameters	β_{10}	β_{11}	β_{20}	β_{21}	$\delta_{1,2}$	$\delta_{2,2}$	σ_{11}	σ_{12}	σ_{22}
Values	−0.5	1	1	−0.5	1.2	0.7	1	−0.8	1
Mean of estimates	−0.494	0.992	1.004	−0.505	1.203	0.703	1.003	−0.802	1.003
Stand. dev. of estimates	0.149	0.141	0.116	0.067	0.097	0.067	0.166	0.181	0.170
Mean of bootstrap stand. errors	0.166	0.148	0.118	0.084	0.087	0.068	0.160	0.173	0.161

Note that due to the re-parametrization we estimate the differences in thresholds rather than the thresholds themselves, but they coincide in the case of only three levels of the categorical variables. In the simulation the averages of the estimated parameters are equal within two significant digits after the decimal point to the parameter values from which the samples were generated except for two parameters. Even these two estimates differ from the true values by < 0.01.

The approximate equality of the standard deviations of the estimates and the bootstrap standard errors confirms that the algorithm is converging as expected. However, larger simulation study that varies the parameter settings is necessary to confirm the above observations.

3.1 Implementation of the Algorithms

For the implementation of the algorithms we created functions in the free software environment for statistical computing and graphics **R** [1]. The **R** code for fitting the presented models is available from the authors.

We want to point out several technical details regarding the implementation of the considered ECM algorithms. In the package **mvtnorm** [17] there are functions for analytical finding of the first two moments of multivariate truncated normal distribution based on the work of Manjunath and Wilhelm [18]. There are also functions for generating random numbers using Gibbs sampling [19] which allows stochastic approximation of the first two moments of the truncated normal distribution. Either can be used for estimation but when the multiple outcomes are only two the analytical calculation is more precise and at least as fast as the stochastic approximation, so we recommend it.

4 Conclusions

In this paper we reviewed EM algorithms for correlated probit models for one longitudinal ordinal outcome and for multiple ordinal outcomes. We studied the performance of the ECM algorithm for multiple ordinal outcomes with a simulation study. Our approach has advantages over alternative estimation methods in that it can handle a large dimension of the multivariate outcome, it can be easily extended to any combination of binary, ordinal and continuous outcomes and it provides asymptotically unbiased estimates. It is also easily implemented in the open-source software environment **R**. Using free software is a premise for wider usage and quicker improvement of the code.

There are several possible directions in which the implementation of the algorithm can be improved. There is a possible extension, called parameter expanded ECM algorithm [20] that can accelerate the speed of convergence of the algorithm. Rather than restrict some parameters (e.g. the variance of the error term) for parameter identifiability up front, this extension allows estimation of all parameters free of restrictions and at the last iteration calculates fully identifiable functions of the parameters (e.g. the ratios of the regression parameters and the squared root of the variance of the errors estimates). An example of implementation of this algorithm can be found in Gueorguieva and Agresti [21]. Another extensions are EM algorithms for MLE of correlated probit model for multiple longitudinal ordinal outcomes and correlated probit model for multiple ordinal outcomes with correlated errors.

Acknowledgements

This work was supported by the European Social Fund through the Human Resource Development Operational Programme under contract BG051PO001-3.3.06-0052 (2012/2014).

References

1. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. (2013) ISBN 3-900051-07-0
2. Bliss, C.I.: The method of probits. Science (1934) 38–39
3. Bliss, C.I.: The method of probits – a correction. Science (1934) 409–410
4. Gueorguieva, R.V.: Correlated probit model. In: Encyclopedia of Biopharmaceutical Statistics. (2006) 355–362
5. Fahrmeir, L., Tutz, G.: Multivariate Statistical Modelling Based on Generalized Linear Models. Second edn. Springer-Verlag, New York (2001)
6. Liu, Q., Pierce, D.A.: A note on Gauss-Hermite quadrature. Biometrika **81**(3) (1994) 624–629
7. Breslow, N.E., Clayton, D.G.: Approximate inference in generalized linear mixed models. Journal of the American Statistical Association **88** (1993) 9–25
8. Wolfinger, R., O'Connell, M.: Generalized linear mixed models: A pseudo-likelihood approach. Journal of Statistical Computation and Simulation **48** (1993) 233–243
9. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society: Series B (Methodological) **39**(2) (1977) 1–22
10. Meng, X.L., Rubin, D.B.: Maximum likelihood estimation via the ECM algorithm: A general framework. Biometrika **80**(2) (1993) 267–278
11. Ruud, P.A.: Extensions of estimation methods using the EM algorithm. Journal of Econometrics **49**(3) (1991) 305–341
12. Kawakatsu, H., Largey, A.G.: EM algorithms for ordered probit models with endogenous regressors. Econometrics Journal **12** (2009) 164–186
13. Chan, J., Kuk, A.: Maximum likelihood estimation for probit-linear mixed models with correlated random effects. Biometrics **53** (1997) 86–97
14. Grigorova, D., Gueorguieva, R.: Implementation of the EM algorithm for maximum likelihood estimation of a random effects model for one longitudinal ordinal outcome. Pliska Stud. Math. Bulgar. **22** (2013) 41–56
15. McLachlan, G.J., Krishnan, T.: The EM Algorithm and Extensions (Wiley Series in Probability and Statistics). 2 edn. Wiley-Interscience (March 2008)
16. Efron, B., Tibshirani, R.J.: An Introduction to the Bootstrap. 1 edn. Volume 57 of Monographs on Statistics & Applied Probability. Chapman & Hall, New York (1994)
17. Wilhelm, S., Manjunath, B.G.: tmvtnorm: Truncated Multivariate Normal and Student t Distribution. (2012) R package version 1.4-7.
18. Manjunath, B.G., Wilhelm, S.: Moments calculation for the double truncated multivariate normal density. `http://ssrn.com/abstract=1472153` (September 11 2009)
19. Wilhelm, S.: Gibbs sampler for the truncated multivariate normal distribution. Electronic (April 6 2012) `http://cran.r-project.org/web/packages/tmvtnorm/vignettes/GibbsSampler.pdf`
20. Liu, C., Rubin, D., Wu, Y.: Parameter expansion to accelerate EM: The PX-EM algorithm. Biometrika **85**(4) (1998) 755–770
21. Gueorguieva, R.V., Agresti, A.: A correlated probit model for joint modeling of clustered binary and continuous responses. Journal of the American Statistical Association **96** (2001) 1102–1112

On a Poisson Negative Binomial Process

Krasimira Y. Kostadinova[1,2]

[1] Faculty of Mathematics and Informatics, Shumen University “K. Preslavsky”,
115, Universitetska St, 9700 Shumen, Bulgaria
kostadinova@shu-bg.net
[2] Faculty of Mathematics and Informatics, Sofia University “St. Kliment Ohridski”,
5, James Bourchier Blvd., 1164 Sofia, Bulgaria

Abstract. In this paper we define a Poisson negative binomial process as a compound Poisson process with negative binomial compounding distribution. We investigate some of its basic properties, recursion formulas and probability mass function. We then describe the defined process as a compound birth process. Then we consider a risk model in which the claim counting process is the defined Poisson negative binomial process. For the defined risk model we derive the joint distribution of the time to ruin and the deficit at ruin as well as the ruin probability. We discuss in detail the particular case of exponentially distributed claims.

Keywords: negative binomial distribution, birth process, ruin probability

1 Introduction

The simplest counting process with various applications is the homogeneous Poisson process. The application of the Poisson process in classical risk models and ruin probability was analyzed for example by Grandell [4, 5] and Rolski et al. [13]. At the same time, different type of generalizations of the Poisson process are given in the literature. Here we can mention, for example the Pólya–Aeppli process, defined by Minkova [11] and characterized by Chukova and Minkova [1]. The Pólya–Aeppli process of order k as a compound Poisson process is defined by Chukova and Minkova [2]. Some modified birth processes are given in Minkova [9].

In Section 2 we define the Poisson negative binomial process as a compound Poisson process with negative binomial compounding distribution. In Section 3 we consider this process as a pure birth process. In Section 4 we consider the Poisson negative binomial risk model and derive a differential equation for the joint distribution of the time to ruin and the deficit at the time of ruin and an expression for the ruin probability. The particular case of exponentially distributed claims is given in Section 5.

2 Poisson Negative Binomial Process

In this section we introduce the Poisson negative binomial process as a compound Poisson process and later we give a second definition for this process as a compound birth process.

We consider the stochastic process $N(t)$, $t > 0$ defined on a fixed probability space $(\Omega, \mathcal{F}, P)$ and given by

$$N(t) = \begin{cases} X_1 + X_2 + \cdots + X_{N_1(t)}, & N_1(t) > 0, \\ 0, & N_1(t) = 0, \end{cases} \tag{1}$$

where X_i, $i = 1, 2, \dots$, are independent, identically distributed (iid) as X random variables, independent of $N_1(t)$. We suppose that the counting process $N_1(t)$ is a Poisson process with intensity $\lambda > 0$ ($N_1(t) \sim Po(\lambda t)$). In this case $N(t)$ is a compound Poisson process. The compounded process $N_1(t)$ is characterized by the following probability mass function (PMF)

$$P(N_1(t) = i) = \frac{(\lambda t)^i e^{-\lambda t}}{i!}, \quad i = 0, 1, \dots \tag{2}$$

and the probability generating function (PGF) is given by

$$\psi_{N_1(t)}(s) = E(s^{N_1(t)}) = e^{-\lambda t(1-s)}. \tag{3}$$

Suppose that the compounding random variable X has a negative binomial distribution, $X \sim NB(r, \gamma)$, with PMF

$$q_i = P(X = i) = \binom{r+i-1}{i} \gamma^r (1-\gamma)^i, \quad i = 0, 1, \dots \tag{4}$$

and PGF

$$\psi_1(s) = \psi_X(s) = \left(\frac{\gamma}{1-(1-\gamma)s} \right)^r. \tag{5}$$

For the PGF of the process $N(t)$, given in (1) we get

$$\psi(s) = \psi_{N(t)}(s) = e^{-\lambda t(1-\psi_1(s))}, \tag{6}$$

where $\psi_1(s)$ is the PGF of the compounding distribution, given by (5).

Definition 1. *The stochastic process, defined by the PGF* (6) *and compounding distribution, given in* (4) *and* (5) *is called a Poisson negative binomial process with notation* $N(t) \sim PoNB(\lambda t, r, \gamma)$.

Remark 1. In the case $r = 1$, the distribution of X in (4) reduces to the geometric distribution and the $PoNB(\lambda t, 1, \gamma)$ distribution coincides with the Pólya–Aeppli distribution, see Johnson et al. [6] and Minkova [10]. The corresponding Pólya–Aeppli process was introduced by Minkova [11] and characterized by Chukova and Minkova [1].

2.1 The Probability Mass Function

The probability function of $N(t)$ is given by expanding the PGF $\psi(s)$ in powers of s. Denote by $P_i(t) = P(N(t) = i)$, $i = 0, 1, 2, \ldots$, the probability mass function of $N(t)$. We rewrite the PGF of (6) in the form

$$\psi(s) = e^{-\lambda t} \sum_{m=0}^{\infty} \frac{(\lambda t)^m}{m!} \psi_1^m(s) = e^{-\lambda t} \sum_{m=0}^{\infty} \frac{(\lambda t \gamma^r)^m}{m!} \frac{1}{(1-(1-\gamma)s)^{rm}}. \quad (7)$$

Denote by $\psi^{(i)}(s) = \dfrac{\partial^{(i)}\psi(s)}{\partial s^i}$ for $i = 0, 1, \ldots$, the derivatives of $\psi(s)$. From (7) we get

$$\psi^{(i)}(s) = e^{-\lambda t}(1-\gamma)^i \sum_{m=1}^{\infty} \frac{(\lambda t \gamma^r)^m}{m!} \frac{rm(rm+1)\cdots(rm+i-1)}{(1-(1-\gamma)s)^{rm+i}}. \quad (8)$$

From Johnson et al. [6], it is known that

$$P_i(t) = \left.\frac{\psi^{(i)}(s)}{i!}\right|_{s=0}. \quad (9)$$

Theorem 1. *The probability mass function of $N(t)$ is given by*

$$\begin{aligned} &P_0(t) = e^{-\lambda t(1-\gamma^r)}, \\ &P_i(t) = (1-\gamma)^i \sum_{m=1}^{\infty} \binom{rm+i-1}{i} \frac{(\lambda t \gamma^r)^m}{m!} e^{-\lambda t}, \quad i = 0, 1, \ldots . \end{aligned} \quad (10)$$

Proof. The initial value $P_0(t) = e^{-\lambda t(1-\gamma^r)}$ follows from the PGF in (6), $\psi(0) = P_0(t)$. Then (10) follows from (8) and (9). □

The following proposition gives an extension of the Panjer recursion formulas (see Panjer [12]).

Proposition 1. *The PMF of the Poisson negative binomial process satisfies the following recursions:*

$$iP_i(t) = (1-\gamma)[(i-1+\lambda t r q_0)P_{i-1}(t) + \lambda t r \sum_{j=1}^{i-1} q_j P_{i-j-1}(t)], \quad i = 2, 3, \ldots, \quad (11)$$

where q_j, $j = 0, 1, \ldots$, are the probabilities of the compounding $NB(r, \gamma)$ random variable, given in (4) *and $P_1(t) = \lambda t r(1-\gamma)q_0 P_0(t)$ with $P_0(t) = e^{-\lambda t(1-\gamma^r)}$.*

Proof. Differentiation in (6) leads to

$$\frac{\partial \psi_{N(t)}(s)}{\partial s} = \frac{\lambda t r(1-\gamma)}{1-(1-\gamma)s} \psi_1(s) \psi_{N(t)}(s), \quad (12)$$

where

$$\psi_{N(t)}(s) = \sum_{i=0}^{\infty} P_i(t)s^i, \quad \frac{\partial \psi_{N(t)}(s)}{\partial s} = \sum_{i=0}^{\infty}(i+1)P_{i+1}(t)s^i, \quad \psi_1(s) = \sum_{j=0}^{\infty} q_j s^j.$$

Equation (12) has the form

$$\begin{aligned}[1-(1-\gamma)s]\sum_{i=0}^{\infty}(i+1)P_{i+1}(t)s^i &= \lambda t(1-\gamma)r\sum_{j=0}^{\infty} q_j s^j \sum_{i=0}^{\infty} P_i(t)s^i \\ &= \lambda t(1-\gamma)r\sum_{i=0}^{\infty}\sum_{j=0}^{i} q_j P_{i-j}(t)s^i.\end{aligned}$$

The recursions (11) are obtained by equating the coefficients of s^i on both sides for fixed $i = 0, 1, 2, \ldots$. □

Remark 2. The mean and the variance of the Poisson negative binomial process are given by

$$\begin{aligned}E(N(t)) &= \frac{1-\gamma}{\gamma}\lambda tr, \\ \mathrm{Var}(N(t)) &= \frac{(1-\gamma)(1+r-\gamma r)}{\gamma^2}\lambda tr.\end{aligned}$$

For the Fisher index of dispersion we obtain

$$\mathrm{FI} = \frac{\mathrm{Var}(N(t))}{E(N(t))} = 1 + \frac{(1-\gamma)(r+1)}{\gamma} > 1,$$

i.e. the Poisson negative binomial process is overdispersed related to the Poisson process.

3 Poisson Negative Binomial Process as a Pure Birth Process

The counting process $\{N(t), t \geq 0\}$ represents the number of times a certain event occurs in time interval $(0, t]$. The transition probabilities of $N(t)$ are specified by the following postulates:

$$P(N(t+h) = n \,|\, N(t) = m) = \begin{cases} 1 - \lambda(1-\gamma^r)h + o(h), & n = m, \\ \lambda q_i h + o(h), & n = m+i, \ i = 1, 2, \ldots, \end{cases} \tag{13}$$

for every $m = 0, 1, \ldots,$ where $o(h) \to 0$ as $h \to 0$. Then, the above postulates yield the following Kolmogorov forward equations:

$$\begin{aligned}P_0'(t) &= -\lambda(1-\gamma^r)P_0(t), \\ P_m'(t) &= -\lambda(1-\gamma^r)P_m(t) + \lambda\sum_{i=1}^{m} q_i P_{m-i}(t), \ m = 1, 2, \ldots\end{aligned} \tag{14}$$

with initial conditions

$$P_0(0) = 1 \quad \text{and} \quad P_m(0) = 0, \quad m = 1, 2, \dots \ . \tag{15}$$

Multiplying mth equation of (14) by s^m and summing for all $m = 0, 1, 2, \dots$, we get the following differential equation for the PGF

$$\frac{\partial \psi_{N(t)}(s)}{\partial t} = -\lambda[1 - \psi_1(s)]\psi_{N(t)}(s). \tag{16}$$

The solution of (16) with the initial condition $\psi_{N(t)}(1) = 1$ is given by

$$\psi_{N(t)}(s) = e^{-\lambda t(1-\psi_1(s))}, \tag{17}$$

where $\psi_1(s)$ is the PGF of the negative binomial distribution, given by (5). It is clear that the PGF (17) identifies the $PoNB(\lambda t, r, \gamma)$ process. This leads to the second definition.

Definition 2. *The counting process $N(t)$ defined by differential equations* (14) *and initial conditions* (15) *is called a Poisson negative binomial process.*

4 Application to Risk Theory

In this section we will apply the Poisson negative binomial process as a counting process in risk model. Consider the standard risk model $\{X(t), t \geq 0\}$, defined on the complete probability space $(\Omega, \mathcal{F}, P)$ and given by

$$X(t) = ct - \sum_{i=1}^{N(t)} Z_i, \quad \left(\sum_{1}^{0} = 0\right). \tag{18}$$

The positive real constant c in this model represents the risk premium rate. The non-negative iid random variables $\{Z_i\}_{i=1}^{\infty}$ are independent of the counting process $N(t)$, $t \geq 0$, and represent the successive claim sizes. Suppose that the claim sizes are distributed as the random variable Z with distribution function F, $F(0) = 0$ and mean value μ. We suppose also that $N(t)$ in the risk model (18), is a Poisson negative binomial process and will call this process Poisson negative binomial risk model. The interpretation of the counting process is the following. If the insurance policies are separated in independent groups, then the number of groups has a Poisson distribution. We suppose that the groups are homogeneous, identically distributed. The number of policies in each of the groups has a negative binomial distribution.

The relative safety loading θ for the risk model in (18) is given by

$$\theta = \frac{EX(t)}{E\sum_{i=1}^{N(t)} Z_i} = \frac{c\gamma}{\lambda \mu r(1-\gamma)} - 1.$$

In the case of positive safety loading $\theta > 0$, the premium income per unit time c should satisfy the following inequality

$$c > \frac{\lambda \mu r(1-\gamma)}{\gamma}.$$

Let $\tau = \inf\{t : X(t) < -u\}$ with the convention of $\inf \emptyset = \infty$ be the time to ruin of an insurance company having initial capital $u \geq 0$. We denote by $\Psi(u) = P(\tau < \infty)$ the ruin probability and by $\Phi(u) = 1 - \Psi(u)$ the non-ruin probability. Let $G(u, y)$ be the joint probability distribution of the time to ruin τ and the deficit at the time of ruin $D = |u + X(t)|$, i.e.

$$G(u, y) = P(\tau < \infty, D \leq y), \quad y \geq 0 \tag{19}$$

and

$$\lim_{y \to \infty} G(u, y) = \Psi(u). \tag{20}$$

The function $G(u, y)$ is defined by Gerber et al. [3] and is analyzed for many of the known risk models, see also Klugman et al. [7], Kostadinova and Minkova [8], Chukova and Minkova [2]. Using the postulates (13) we get

$$\begin{aligned} G(u,y) &= (1-\lambda(1-\gamma^r)h)G(u+ch,y) \\ &+ \lambda \sum_{i=1}^{\infty} q_i h \left[\int_0^{u+ch} G(u+ch-x,y)\, dF^{*i}(x) + F^{*i}(u+ch+y) - F^{*i}(u+ch) \right] \\ &+ o(h), \end{aligned}$$

where $F^{*i}(x)$, $i = 1, 2, \ldots$, is the distribution function of $Z_1 + Z_2 + \cdots + Z_i$, and the probabilities q_i, $i = 0, 1, \ldots$, are given by (4). Rearranging the terms leads to

$$\begin{aligned} \frac{G(u+ch,y) - G(u,y)}{ch} &= \frac{\lambda(1-\gamma^r)}{c} G(u+ch,y) \\ &- \frac{\lambda}{c} \sum_{i=1}^{\infty} q_i \left[\int_0^{u+ch} G(u+ch-x,y)\, dF^{*i}(x) + F^{*i}(u+ch+y) - F^{*i}(u+ch) \right] \\ &+ \frac{o(h)}{h}. \end{aligned}$$

Let

$$H(x) = \sum_{i=1}^{\infty} q_i F^{*i}(x)$$

be the defective probability distribution function of the aggregate claims with

$$H(0) = 0 \quad \text{and} \quad H(\infty) = 1 - \gamma^r.$$

By letting $h \to 0$ we obtain the following differential equation

$$\frac{\partial G(u,y)}{\partial u} = \frac{\lambda(1-\gamma^r)}{c} G(u,y) - \frac{\lambda}{c}\left[\int_0^u G(u-x,y)\,dH(x) + [H(u+y) - H(u)]\right]. \quad (21)$$

In terms of the proper probability distribution function $H_1(x) = \dfrac{H(x)}{1-\gamma^r}$ equation (21) can be rewritten by

$$\frac{\partial G(u,y)}{\partial u} = \frac{\lambda(1-\gamma^r)}{c}\left[G(u,y) - \int_0^u G(u-x,y)\,dH_1(x) - [H_1(u+y) - H_1(u)]\right]. \quad (22)$$

Theorem 2. *The function $G(0,y)$ for the defined risk model is given by*

$$G(0,y) = \frac{\lambda(1-\gamma^r)}{c}\int_0^y [1 - H_1(u)]\,du. \quad (23)$$

Proof. Integrating (22) from 0 to ∞ with $G(\infty, y) = 0$ leads to

$$- G(0,y) = \frac{\lambda(1-\gamma^r)}{c}\left[\int_0^\infty G(u,y)\,du - \int_0^\infty \int_0^u G(u-x,y)\,dH_1(x)\,du - \int_0^\infty (H_1(u+y) - H_1(u))\,du\right]. \quad (24)$$

The change of variables in the double integral and simple calculations yield

$$G(0,y) = \frac{\lambda(1-\gamma^r)}{c}\int_0^\infty [H_1(u+y) - H_1(u)]\,du$$

and then (23) holds. □

4.1 Ruin probability

Theorem 3. *For $u \geq 0$, the probability of ruin $\Psi(u)$ and the non-ruin probability $\Phi(u)$ satisfies the equations:*

$$\frac{\partial \Psi(u)}{\partial u} = \frac{\lambda(1-\gamma^r)}{c}\left[\Psi(u) - \int_0^u \Psi(u-x)\,dH_1(x) - [1 - H_1(u)]\right], \quad (25)$$

$$\frac{\partial \Phi(u)}{\partial u} = \frac{\lambda(1-\gamma^r)}{c}\left[\Phi(u) - \int_0^u \Phi(u-x)\,dH_1(x)\right]. \quad (26)$$

Proof. The result (25) follows from (20) and (22). Then, applying the condition $\Phi(u) = 1 - \Psi(u)$, we obtain (26). □

Theorem 4. *The ruin probability with no initial capital satisfies*

$$\varPsi(0) = \frac{r(1-\gamma)}{c\gamma}\lambda\mu. \tag{27}$$

Proof. According to (20) and (23) we have

$$\varPsi(0) = \lim_{y\to\infty} G(0,y) = \frac{\lambda(1-\gamma^r)}{c}\int_0^\infty [1-H_1(u)]\,du.$$

Let X be a random variable with distribution function $H_1(x)$. By the definition of $H_1(x)$ and $EZ = \mu$ we obtain

$$EX = \frac{\mu}{1-\gamma^r}[q_1 + 2q_2 + 3q_3 + \cdots] = \frac{\mu}{1-\gamma^r}\frac{r(1-\gamma)}{\gamma}.$$

Using the fact that $EX = \int_0^\infty [1-H_1(x)]\,dx$ we obtain (27). □

Remark 3. Based on (27), it is easy to see that the ruin probability with no initial capital does not depend on t.

5 Exponentially Distributed Claims

Let us consider the case of exponentially distributed claim sizes, i.e. $F(u) = 1 - e^{-u/\mu}$, $u \geq 0$, $\mu > 0$. The function $H(x)$ in this case has the form

$$H(x) = \gamma^r \sum_{i=1}^{\infty} \binom{r+i-1}{i} (1-\gamma)^i F^{*i}(x),$$

where

$$F^{*i}(x) = 1 - \sum_{j=0}^{i-1} \frac{(x/\mu)^j}{j!} e^{-x/\mu}.$$

For the proper distribution function $H_1(x)$ we obtain

$$H_1(x) = 1 - \frac{\gamma^r}{1-\gamma^r} \sum_{i=1}^{\infty} \binom{r+i-1}{i} (1-\gamma)^i \sum_{j=0}^{i-1} \frac{(x/\mu)^j}{j!} e^{-x/\mu}. \tag{28}$$

Substituting (28) in (23) we obtain the initial condition in the case of exponentially distributed claims

$$G(0,y) = \frac{\lambda\mu\gamma^r}{c} \sum_{i=1}^{\infty} \binom{r+i-1}{i} (1-\gamma)^i \sum_{j=0}^{i-1} \frac{\gamma\,(y/\mu, j+1)}{\varGamma(j+1)},$$

where $\gamma(x,\alpha) = \int_0^x t^{\alpha-1}e^{-t}\,dt$ is the incomplete Gamma function and $\varGamma(j+1)$ is the Gamma function.

Acknowledgements

The author is thankful to the anonymous reviewers and the editor for making some useful comments and suggestions. This work was supported by the European Social Fund through the Human Resource Development Operational Programme under contract BG051PO001-3.3.06-0052 (2012/2014) and by grant RD-08-255/2013 of Shumen University, Bulgaria.

References

1. Chukova, S., Minkova, L.D.: Characterization of the Pólya–Aeppli process. Stochastic Analysis and Applications **31** (2013) 590–599
2. Chukova, S., Minkova, L.D.: Pólya–Aeppli of order k Risk Model. Commun. Statist. – Simulation and Computation (to appear)
3. Gerber, H., Goovaerts, M., Kaas, R.: On the probability and severity of ruin. ASTIN Bull. **17** (1987) 151–163
4. Grandell, J.: Aspects of Risk Theory. Springer-Verlag, New York (1991)
5. Grandell, J.: Mixed Poisson Processes. Chapman & Hall, London (1997)
6. Johnson, N.L., Kemp, A.W., Kotz, S.: Univariate Discrete Distributions. Wiley Series in Probability and Mathematical Statistics, 3th ed. (2005)
7. Klugman, S.A., Panjer, H., Willmot, G.: Loss Models: From Data to Decisions. John Wiley & Sons, Inc. (1998)
8. Kostadinova, K., Minkova, L.D.: On the Poisson process of order k. Pliska Stud. Math. Bulgar. **22** (2013) 117–128
9. Minkova, L.D.: Inflated-parameter modification of the pure birth process. C. R. Acad. Bulg. Sci. **54(11)** (2001) 17–22
10. Minkova, L.D.: A Generalization of the Classical Discrete Distributions. Commun. Statist. - Theory and Methods **31** (2002) 871–888
11. Minkova, L.D.: The Pólya-Aeppli process and ruin problems. J. Appl. Math. Stoch. Anal. **2004**(3) (2004) 221–234
12. Panjer, H.: Recursive evaluation of a family of compound discrete distributions. ASTIN Bull. **12**(1) (1981) 22–26
13. Rolski, T., Schmidli, H., Schmidt, V., Teugels, J.: Stochastic Processes for Insurance and Finance. John Wiley & Sons, Chichester (1999)

Modelling of the Amount of Chlorophyll-*a* Contained in the Phytoplankton Population by Branching Process

Antoanela Terzieva

Faculty of Mathematics and Informatics, Sofia University "St. Kliment Ohridski",
5, James Bourchier Blvd., 1164 Sofia, Bulgaria
Antoanela.t@abv.bg

Abstract. This research focuses on the phytoplankton community and the concentration of chlorophyll-a, which is contained in the phytoplankton cells. We investigate a phytoplankton population and derive a mathematical model. We model and explore a branching process (BP) $(Z_t : t \geq 0)$ and focus on the case where such process may be treated as one-type Bellman–Harris branching processes (BHBP). There is a correlation between the amount of phytoplankton's cells and the amount of chlorophyll contained within. Therefore we propose the same branching process model to be applied to the amount of chlorophyll-a. Instead of a cell the role of the particle will be played by the amount of chlorophyll-a, which the cell contains. Concrete results are obtained for the correlation. Using this model, we will be able to predict the future behavior of the chlorophyll-a concentration and, hence, the phytoplankton population.

Keywords: chlorophyll-a, phytoplankton population, branching process

1 Introduction

It is customary to call as phytoplankton the microscopic unicellular plant species in aquatic environment. Phytoplankton's cells can divide or die. From now on we mean death for disappearance of the cell from the population by means other than division. Under growth rate is meant the rate of increase in size per unit time. Let us assume that the age and volume uniquely determine the growth rate of a cell and its probability of division per unit time. According to the theory [2, 6] we have to assume exponential growth rate. On the one hand, in [10], the authors propose a one-dimensional model that describes the aggregation behavior of phytoplankton on the vertical component of water column; on the other hand, they perform mathematical analysis of this model and explore its asymptotic behavior. The authors in [5] offer intraspecific competition of picophytoplankton for light and nutrients. In [4] the authors consider the two species competitive delay plankton allelopathy stimulatory model system. They show the existence and uniqueness of the solution of the deterministic model. We define the random variables τ_i, ξ, δ_i, m_i, $i = 1, 2, \ldots, n$, over the probability space $(\Omega, F, \mathbb{P})$. Let the numbers λ_l^i, μ_l^i are positive and real.

In [1] the authors consider the three-species nutrient-phytoplankton-zooplankton model and deal with a nutrient-plankton model in an aquatic environment in the context of phytoplankton bloom. Toxin producing phytoplanktons are assumed to play the key role.

2 General Remarks and Simplification

It is well known that the number of offspring is exactly two, and both are born at the same time, but they can belong to different types. After a random time the progenitor splits into two offspring. We study a branching stochastic process $(Z_t : t \geq 0)$ having the non-negative integers as state-space. The process can be thought of as representing an evolving population of particles. The number of particles in the population at time t is denoted by a random variable Z_t. It starts at time $t_0 = 0$ with Z_0 particles. The considered by us phytoplankton population contains many species, especially phytoplankton in the Bulgarian Black Sea Coast contains about 600 different species. For each phytoplankton species there is not just one cell, but a huge amount of cells. We assume that there is a total of n species for any positive integer n and the defined above process Z_t starts with one piece of each type because another cases can reduce to this. We start with one participant from each species, i.e. with one new-born particle T_i from any i-th specie, $i = 1, 2, \ldots, n$, i.e. $Z_0 = (1, 1, \ldots, 1)$. After a random time the progenitor splits into two offspring with probability $\delta_i(t)$ or dies without giving offspring with probability $1 - \delta_i(t)$, $i = 1, \ldots, n$. The process $(Z_t : t \geq 0)$ is volume-structured with minimum volume for splitting. We can construct a branching process starting with a rooted binary tree T and a collection $(\xi(k))$ of independent copies of a random variable ξ over the probability space $(\Omega, F, \mathbb{P})$. There is assumed the additional structure of particle's growth rate motion on the line. We interpret the growth rate as a random variable, which move on the line. Consider a parent particle at the point x_0 assuming a growth rate $r_0 = x_0$. At its death it splits into two particles with probability δ_i, which growth rates then move to the random points $r_{01} = x_0 + X$ and $r_{02} = x_0 + Y$. Subdivision of subtypes of the particles T_i for any fixed number i is based on the growth rate. Let us call the amount of chlorophyll-a, which the cell contains, a *chlorophyll unit.* All participating cells (or chlorophyll units) can have similar, but different growth rates, and from there respectively all particles will be of a different type. The problem is to model the dynamics of the phytoplankton's population or accordingly chlorophyll' units population by a branching process. We have a stochastic process in which each participating particle is from its own unique type.

According [6] we will consider that individual growth is allays exponential, but there are some variations between the growth rate of each child and the mother. Different types of phytoplankton will determine respectively the different units chlorophyll-a, i.e. we will consider the particles of type T_i, $i = 1, 2, \ldots, n$. From now on, when there is no danger of misunderstanding, we will talk about the particle T_i, given that it may represent both cell of phytoplankton and chlorophyll unit of that cell. All particles T_i have an exponential distribution function of growth, but with different parameters. They depend on the value of the parameter in the exponential distributed life length by the mother, but may differ. For the particle of type T_i, let us call shortly "T_i". We will take into account that [6]: Individual growth is exponential with growth rate that consists of two parts: a latent factor handed down by the mother, which represents inheritance, and an individual contribution.

The focus of our study falls on grand total, so that fluctuations from the mean for each type can be ignored. Therefore we assume the cells of each species are divided in half after reaching a fixed average sizes B_i, $i = 1, 2, \dots, n$, for any species and then they grow by the exponentially distributions.

From what we said above and taking into account [6], we can write our assumption for any T_i:

2.1 Conditions

1. Length of life is exponential distributed random variable.
2. The type is the growth rate; it depend on birth quantity, but not completely. The growth rate consists of two parts: a latent factor, and an individual contribution.
3. The particle splits into two daughter particles of precisely equal volumes.
4. There exists a critical size α such that all particles are born in volume smaller than α, and then grow past α before division.

3 Model Description

For the sake of brevity, we use i.i.d. for independent and identically distributed. Practically there are always a lot of cells at the beginning. From now on, let us focus only on the behaviour of one of the six hundred types of particles corresponding to different species of phytoplankton. A particle T_i of any fixed species grows exponentially and has limited exponential growth rate $\lambda^i \in [a_i, b_i]$. We divide the interval $[a_i, b_i]$ into θ_i number of subintervals. Depending on which one of the subintervals falls within the growth rate, we consider θ_i types of particles: H_l^i, $l = 1, \dots, \theta_i$. Each H_l^i has fixed λ_l^i. We will write that

$$\lambda_l^i := a_i + \left(l - \frac{1}{2}\right) \frac{b_i - a_i}{\theta_i}.$$

If we make the number $\varDelta_i = \dfrac{b_i - a_i}{\theta_i}$, then λ_l^i falls right in the middle of the l-th subinterval.

We assume that each particle T_i has a lifetime random variable τ_i with an exponential distribution. At the end of its life it is divided into two equal parts with similar but different growth rates or dies without splitting. Depending on the subintervals in which those growth rates fall, we consider θ_i number of types of particles, all of them as with growth rates and the length of life distributed exponentially and having respectively θ_i different growth rates $\lambda_1, \lambda_2, \dots, \lambda_{\theta_i}$. In general, offspring and parent are of different types, and the children are from different types.

Taking in to account conditions 2.1 we will model the T_i-population by multi-type age-dependent branching process W_t as follows.

Let $p^i_{l,j}$ be the probability, that a daughter of H^i_l is H^i_j. For example, we can assume normal or uniformly distributed shift Δ_i of the exponential growth rates over the interval $[a_i, b_i]$.

We define the random variables τ_l, ξ and m_l on the probability space $(\Omega, F, \mathbb{P})$, $l = 1, 2, \dots, \theta_i$. The particles of W_t can be from θ_i different types. They evolve with lifetimes $\tau^i_l \in \mathrm{Exp}(\mu^i_l)$, $l = 1, \dots, \theta_i$, and distribution functions $G^i_l = \mathbb{P}\{\tau^i_l \leq t\}$. After time τ^i_l of their life, each particle splits, according to the probability law $\{p^i_{l,k}\}$, into two new particles $H^i_1, H^i_2, \dots$, or $H^i_{\theta_i}$ from some of the fixed θ_i types or dies without offspring.

At a given moment of time t_0 of the particular type T_i, there exists a huge number of particles corresponding to the cells. From the total amount of phytoplankton cells, let us isolate the amount corresponding to the i-th type of phytoplankton, i.e. the sum of all T_i. Thus, we obtain the amount of all particles T_i at a given time t_0; it is a random variable, which can be treated as a sum of random variables, divided on the basis growth rate. Let us use the designation § for the amount at a moment t_0. Then

$$\S T_i = \S H^i_1 + \S H^i_2 + \cdots + \S H^i_{\theta_i}.$$

Let us consider an arbitrary time t_0, in which there are many of the $H^i_1, \dots, H^i_{\theta_i}$. According to the properties "loss of memory" of the exponential distribution, we are entitled to assume that all presence particles are born at time t_0, i.e. all are from age zero. Let us denote by H^{ik}_l the k-th generation of the particle type l for $k = 1, 2, 3, \dots$, $l = 1, 2, \dots, \theta_i$. It is clear that k-th generation will amount to 2^k number of particles. Let at time t^l_k the splitting of the H^k_l, $k = 1, 2, \dots$, occurs. We define $t_k := \min[t^1_k, \dots, t^{\theta_i}_k]$ for $k = 1, 2, \dots$. For short, we will use the following symbols for random variables $t_1 - t_0$ and $\lambda_1 + \cdots + \lambda_{\theta_i}$: $Y = Y_1 := t_1 - t_0$, $\lambda^i = \lambda^i_1 + \cdots + \lambda^i_{\theta_i}$.

The following condition is fulfilled: $Y = \min[\tau^i_1, \dots, \tau^i_{\theta_i}]$, $Y \in \mathrm{Exp}(\mu^i)$. Obviously, we can choose sufficiently small number l such that there are available all types of $H_l, \dots, H_{\theta_i}$. Based on Conditions 2.1 we assume some restrictions for the variation of Δ_i.

Similarly, we get $Y = Y_k := t_k - t_{k-1} \in \mathrm{Exp}(\mu^i)$ for $k = 1, 2, \dots$. There is one split in any period of time length Y, $Y \in \mathrm{Exp}(\mu^i)$.

Because the feature "no memory" of the exponential distribution, we can equate the lifetime of all the particles of k-th generation for each $k = 1, 2, \dots$, i.e. the k-th generation is born for all types at time t_k. Let us handle the newborns H_l from all parents k-th generation, i.e. at time t_k, as a particles generated by the particle H^k_l. In this way each particle from the k-th generation is assumed to have life-length random variable τ^i_k and probability distribution $G^i(t) = \mathbb{P}\{\tau^i_k \leq t\}$. At the end of its life, it is transformed into a random number of particles ξ_k of the same type according to the probability law p^i_j. Thus, if we formally swap children we can reduce the multi-type age-dependent branching process W_t to the θ_i in number of the well known single type particles Bellman–Harris branching processes (BHBP) U^l_t, $l = 1, \dots, \theta_i$, with generating function (for the sake of brevity will miss subscripts l):

$$F(t,s) = \mathbb{E}[s^{U_t} | U_0 = 1], \quad t \geq 0.$$

Accordingly the theory (see [8]) it is fulfilled

$$F(t,s) = s(1 - G(t)) + \int_0^t f(F(t-u,s))\, dG(u)$$

for the generating function of U_t.

For $m(t)$ being the measure, $m(t) := \mathbb{E}\xi(t)$, number $\nu > 0$, we consider the equation

$$\nu \int_0^\infty e^{-\alpha u}\, dG(u) = 1 \tag{1}$$

For $A(t) := \mathbb{E}U_t$ we have

$$A(t) = 1 - G(t) + m \int_0^t A(t-u)\, dG(u).$$

If (1) has a solution α then this number is called *Malthusian parameter* for U_t. If $m > 1$ we have

$$A(t) \sim e^{\alpha t} \frac{\displaystyle\int_0^\infty e^{-\alpha u}(1 - G(u))\, du}{\displaystyle m \int_0^\infty u e^{-\alpha u}\, dG(u)} = e^{\alpha t} \frac{m-1}{\displaystyle \alpha m^2 \int_0^\infty u e^{-\alpha u}\, dG(u)}.$$

Using the above general formulas we derive results for the case under consideration.

4 Chlorophyll-*a*

In the simplest case, we assume that there are obtained data for the concentration of chlorophyll-*a* in the phytoplankton community by water-sampling. Such values for the amount of chlorophyll-*a* may be obtained not only by the sampling, but also via satellite. This may be the reason to assume as a single particle in our process, not the whole cell phytoplankton, but only contained in the cell quantity of chlorophyll-*a*. In this case, the correlation coefficient between phytoplankton and in it contained chlorophyll-*a* is known. The phytoplankton community and respectively its content of chlorophyll-*a* grows or reduces in volume. Two kinds of relations (non-allometric and allometric) of chlorophyll with phytoplankton density, biovolume and surface area were investigated for example in [11].

Using all arguments stated by us above, including the suggested mathematical model, allow us to repeat the process that has particulate quantities of chlorophyll from the cells of phytoplankton.

5 Conclusions

We propose a branching stochastic process of evolving populations of particles as a model of phytoplankton community. Each particle represents a single phytoplankton's cell. The idea is that based on the amount of chlorophyll for the practice there are methods for easy reading. We can draw some conclusions about the phytoplankton amount dynamics, considering specific to any particular area number of phytoplankton species and the corresponding correlation coefficients between phytoplankton and chlorophyll-a. Until now, we have never encountered in the literature such model of branching process. The model simplifies to the well-studied BHBP. We are about to test the agreement between the distribution, obtained from the model and experimental data sampled along the Bulgarian Black Sea coast. To do this, we first need to get an accurate estimate of the average quantity of phytoplankton, which feeds marine inhabitants. This must be done depending on the season and the geographical location of phytoplankton's population.

6 Final Notes

The phytoplankton controls the global carbon cycle which has a significant impact on the climate regulation. Phytoplankton cells have the ability of forming dispersed aggregates in the water column,which constitute the main food available to the early larval stages of many fish species, including the anchovy. At such stages, larvae are passive and can only eat the prey passing in very close vicinity [10]. In addition the phytoplankton plays the key role at the base of the marine food chain. Given any mechanisms to influence phytoplankton's population size we could possibly increase fish production for example. Phytoplankton is of great importance in ecological sense, climate regulation and feeding people.

Acknowledgements

This work is partially supported by funds of the National Project for supporting doctoral students and young scientists in Mathematics and Computer Science at Sofia University "St. Kliment Ohridski" and co-financed by the European Social Fund through the Human Resource Development Operational Programme under contract BG051PO001-3.3.06-0052 (2012/2014).

References

1. Mukhopadhyay, B., Bhattacharyya, R.: Modeling phytoplankton allelopathy in a nutrient-plankton model with spatial heterogeneity. Ecological Modeling **198**(1–2) (2006) 163–173
2. Anderson, E.C., Bell, G.I., Petersen, D.F., Tobey, R.A.: Cell Growth and Division: IV. Determination of Volume Growth Rate and Division Probability. Biophysicalical Journal **9** (1969) 246–263

3. Bell, G.I., Anderson, E.C.: Cell Growth and Division: I. A mathematical model with applications to cell volume distribution in mammalian suspension cultures. Biophysicalical Journal **7** (1967) 329–351
4. Abbas, S., Banerjee, M., Hungerbühler, N.: Existence, uniqueness and stability analysis of allelopathic stimulatory phytoplankton model. Journal of Mathematical Analysis and Applications **367** (2010) 249–259
5. Denaro, G., Valenti, D., La Cognata, A., Spagnolo, B., Bonanno, A., Basilone, G., Mazzola, S., Zgozi, S.W., Aronica, S., Brunet, C.: Spatio-temporal behavior of the deep chlorophyll maximum in Mediterranean Sea: Development of a stochastic model for phytoplankton dynamics. Ecological Complexity, **13** (2013) 21–34
6. Haccou, P., Jagers, P., Vatutin, V.A.: Processes Branching: Variation, Growth, and Extinction of Populations. Cambridge University Press, Cambridge, UK (2005)
7. Athreya, N.: Branching processes. Springer (1972)
8. Berard, J., Gouere, J.-B.: Survival probability of the branching random walk killed below a linear boundary. Electronic Journal of Probability **16** (2011) Paper no. 14 396–418. Journal URL: `http://www.math.washington.edu/`
9. Adioui, M., Arino, O., El Saadi, N.: A nonlocal model of phytoplankton aggregation. Nonlinear Analaysis: Real World Applications **6** (2005) 593–607
10. Kalchev, R.K., Beshkova, M.B., Boumbarova, C.S., Tsvetkova, R.L., Sais, D.: Some allometric and non-allometric relationships between chlorophyll-a and abundance variables of phytoplankton. Hydrobiologia **341**(3) (1996) 235–245
11. Kalchev, R.K., Beshkova, M.B., Boumbarova, C.S., Tsvetkova, R.L., Sais, D.: Some allometric and non-allometric relationships between chlorophyll-a and abundance variables of phytoplankton. Hydrobiologia **341**(3) (1996) 235–245

Recommendations in Social Networks: an Extra Feature or an Essential Need

Milen Chechev and Ivan Koychev

Faculty of Mathematics and Informatics, Sofia University "St. Kliment Ohridski",
5, James Bourchier Blvd., 1164 Sofia, Bulgaria
{milen.chechev,koychev}@fmi.uni-sofia.bg

Abstract. This paper analyzes user's need of content recommendation at the social network Facebook. It presents results from a survey on real social network's users. The results shows that Facebook users need better interface for news feed browsing. It have to provide better information filtering options, recommendation system and options for manual refinement of the results from it. The collected information from the survey is used to determine features which an application has to provide as social network news feed browser and to receive user's trust. Further some implementation details and faced difficulties are presented.

Keywords: recommender system, social recommendation, Facebook

1 Introduction

Recommendation systems and social networks are two of the hottest topics at the information society at the beginning at 21st century. According to the statistics internet is used by 2.4 billion people [1] and social network Facebook already have more than a billion users [2, 3]. These statistics indicate the importance of the social networks as irreplaceable part of the online live of the people.

At the beginning Facebook was created as communication environment for university's students, but it quickly expands and nowadays it's used by everyone for photo and information sharing. Users share information about their education, workplace, personal relationships, their activities and opinions on different social topics. Its simple interface and the abilities for information sharing make it attractive and lead its continuous expansion. However the increasing number of users and content at the social network bring some difficulties to the users about managing information sources. According to Robin Dunbar [17] the human have physical limitation of the social connections, which it can maintain. He determines 150 as limit of the maintainable connections per person. The official Facebook statistics [3] shows that the average number of the friends at Facebook for users, which are more than two years at the social network, is 300. This large number of friends cause problems in maintenance of the connections and cause the need of classification on the connections to close friends and acquaintances. Further the users are overloaded with information from their connections and they spend most of their time to filter their social feed and to find information they are interested in. This problem looks similar to the main problem at the information retrieval – finding useful information in a document's set, the success in dealing with it made Google the most popular web search engine. Unlike it at the social networks the users does not know what exactly they are interested in. The information need to be automatically filtered and ranked based on explicit and implicit information about the user's preferences.

2 Related Work

Last decade recommendation systems are widely developed and improved, because of their commercial value for increasing of the purchases at online stores. To stimulate the research activity at this field, competitions on real user's data NetflixPrize [4], ECML-PKDD 2011 [6], HåtRec [14], KDDCup [5, 7] was organized. The series started with the NetflixPrize competition which looks for the best recommendation system for internet site for movie rental. Further a trend of changing the task from recommendation at internet store to recommendation at social network is observed. For example: the HitRec competition exploits data from musical social network – lastfm and internet bookmarks – delicious; KDDCup 2012 data are provided from the most popular Chinese social network Transent Weibo [8].

Recommendations at the social network are based on the relations between the users. There is a research which shows that this relations can be modeled as connection based on trust [9, 15, 16]. The trust at the social network can be two types – explicit and implicit, as an example of trust network with explicit trust is the review's network Epinions [18]. Massa and Avesani [9] study the dissemination of the trust and distrust at the social networks. Their research shows that the recommendation strategies based on trust provide good results even when the information is insufficient for the others recommendation approaches.

Social Network Facebook can be considered as network of trust. The trust is given when user likes publication of another user As this trust can be considered as implicit because the users did not explicitly specify that they trust the author. Some recommendation options are already integrated at Facebook. For browsing the news feed, Facebook users have two views – "Most Recent News" or "Top News". First view shows the news sorted on their publication time. It is comfortable for users which often read their news stream otherwise Facebook suggest the view "Top News". It sort the news based on several factors like how often the user use the social network, what are the relations between the user and the user who has published the news, etc. [10]. The "Top News" view is useful for users which do not spend much time at the social network, but it works as a black box and the users feels uncertain to use it because there is not an explanation how exactly the ranking is made and they cannot refine it based on their preferences.

Our research shows that Facebook users need new and flexible application, which provides similar functionality to the "Top News" view of the user's stream, but with additional options for manual refinement and filtering. Next section presents the user survey that determines the main functionality to be included in a social network recommending application.

3 Survey of User Needs

As part of the experiments to determine necessity of additional application for content recommendations at the social network, an online survey was conducted. Participants at the survey were 114 Facebook users with the following demographic characteristics: 70% of participants are between 18 and 30 years old, and 30% are older than 30 years. Three per cent of the asked users have secondary education, 30% have BSc degree, 60% have MSc degree and 7% are PhD. 45% from the participants are males and 55% are females.

The questionnaire which all participants answered aims to test the following hypothesis:

- There is a necessity of new application that helps Facebook users and recommends interesting news according to their preferences.
- Users are generally skeptical to recommendation systems that work as black boxes. They want an explanation about the recommendations and options for manual tuning of the system.
- To be more attractive a recommendation application has to provide functionality for information filtering and easily browsing of the content.

Users were asked about the time they spend at the social network. The answers shows that 25% of the users spend just several minutes per day, 47% claim that they spend more than a half hour per day and 28% more than an hour per day. Then users were asked series of questions about their concrete activity at the network. The question “What percent of your time at the social network you spend for an activity?” was asked for the different activities at the social network – reading news feed, browsing friends photos, chat with friends, playing games, etc.

Users determine news feed reading as the activity, which is the most time consuming, at their interaction with the social network. Answers show that 62% of the people spend more than a half of their time, at the social network, for reading the news feed. Second place for most time consuming activity is hold from the activity – browsing friend’s photo albums. Seventy per cent from the users claims that they spend more than 30% from their time for that activity. Detailed distribution of the percentages spend by the users for the activities – reading news feed and browsing friends photo albums can be seen in Fig. 1. The statistics shows that most of the users spend less than 20% for all other activities at the social network. Detail distributions on the percentages spend for the activities – chat with friends, publishing content and playing games are presented in Fig. 2.

Next set of questions tried to determine user’s needs and what are the new features they require from the social network. 54% of the users answer that they need better search options at the social network. They want to be able to search

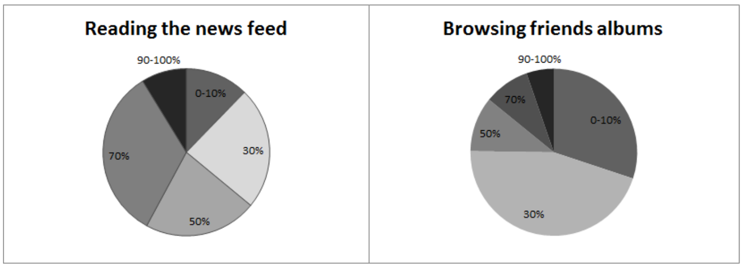


Fig. 1. Main activities at the social network

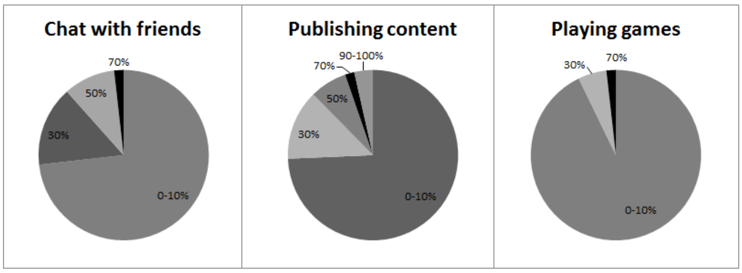


Fig. 2. Additional activities at the social network

for the news at their stream in similar way as they can search at Google. 52% of the users wants to have an options to set manually the priority of the different information sources, as only 35% considered automatically selected priority like something they need. Furthermore, 54% of the users answer that system that automatically create their user profile is acceptable only if there are options for manual refinement of the automatically selected source's weights. Only 25% of the users consider as acceptable the system to continue drift the weights based on their activity after they manually refine their profile. In addition users report high interest on different filtering options – filter read/unread news, filter by date, by user activity, by publication type, etc.

Users report very low activity for adding new applications. Only 13% of the users claim that they have added at least one application to their profiles at the last month. Nevertheless 69% report willingness to add new application if it gives them possibilities for easy filtering of their news feed and options for automatic and manual calibration of the priority of the different information sources.

The results from the survey confirms the preliminary formulate hypothesis. There is a need of new application that helps users in their everyday use of the social network (69% of the requested users answered that they are ready to use such kind of application). The users are generally skeptical to the recommendation systems (only 25% of them recognize as acceptable their priorities to be automatically calibrated by recommendation system) and more than a half of the users (54%) want to be able to change manually their user profile. Despite user's skeptical opinion about automatic recommendation system, the need of better tool for filtering the content at the news feed determine good chances of fast dissemination. Users are ready to use application that support their everyday news feed activity and this gives the application opportunity of collecting information about more users which will cause and better recommendations.

4 Available Information and Its Accessibility

Facebook provides for developers program interface – graph API and FQL queries, which are used to access information about users and their friends, according to strict access policy [11].

Collected data for the purposes of the experiments is in the following categories:

- Objects liked by the user
- Comments by the user
- Shared pages
- Published photos and statuses
- User's news feed
- News published by the user's friends.

The access to the friend's data is restricted and even when the application have access to see what user's friends publish to the social network, it cannot access explicitly what user's friends like. Despite this data can be accessed implicitly from the detailed information about the objects accessible by the application.

There are several difficulties during the process of gathering data from Facebook most of them are due to limitations which Facebook have on its program API. The following limitations are noticed:

- The maximum number of requests to the Facebook API is 600 for the period of 10 min. This limit restricts an application of retrieving detailed information for each accessible object on regular bases.
- The access to the historical data is limited to:
  - Last 100 liked objects per user.
  - News feed for the last 2 days.
  - Maximum 420 publications from the user feed.
  - Without additional query there is an access to maximum 5 likes and comments per content.

Gathered data shows the following statistics:

- There are around 200 new contents which are published at the user's feed per 24 hours. This content collects around 400 likes.
- Most of the examined users have more than 300 friends.
- The execution time for execution of a query to the Facebook API depends on the network connectivity and current workload of the social network, as in the provided experiments 50% of the queries took more than 15 s for response. The time consumption determines the use of asynchronous calls and processing of the results at different threads.

5 Implementation

Multilayer architecture is selected for the new Facebook news feed recommender application. Presentation layer include user interface providing the required functionality and which is called by the user from as an embedded environment at the social network. The application layer contains the business services which gather information from the social network process it and send it to the data layer where it is stored (see Fig. 3).

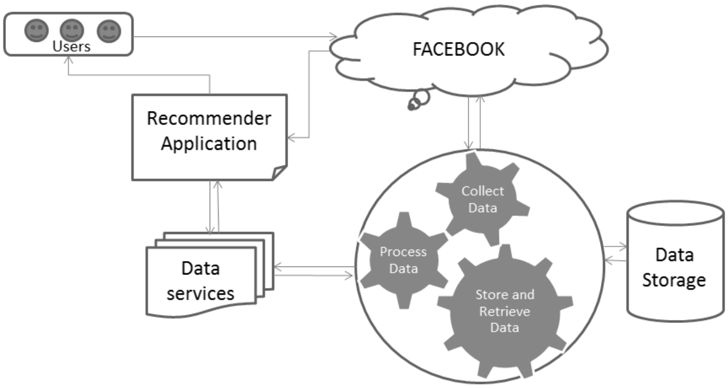


Fig. 3. Architecture of the recommendation application

When user open the application for first time it ask about permission to access user's basic information and news feed. Once the permissions are granted it starts immediately to process the data as the average time that the application needs to provide results to the user is around half a minute. Meanwhile a welcome page is displayed which provide information about the application in textual and video format.

After the data is processed and the application can provide results, the functionality for reading and filtering of the news feed become active. The system offer search functionality and all gathered objects are included at the search results. The user profile browsing and calibration remains inactive until all the needed information is processed.

After the user profile is created it is used for ranking of the news feed and the user receives access to browse and calibrate it. The offered calibration activities are – changing of the preferences to a particular friend, changing the preferences to a particular keyword, adding or removing a keyword. All changing values are stored at the database and are later used by the system to recalculate the user's profile.

Spreading of the application at the social network is crucial for his evaluation. There are two different ways to advertise the application – explicit and implicit. The explicit way is via invitations send to the friends of the users which already used the application. The implicit way is via link to the application when user shares content through the application interface.

6 Recommendation Approaches

The system uses both explicit and implicit information about user's preferences and processes it at three recommendation approaches:

- Collaborative recommendations – it use as source of information user's object marked as "liked" and the recommendation is made based on which users have preferences similar to ours. The selected algorithm for collaborative filtering is Slope One [13]. It takes as input parameter user-object table which contains 1 if the object is liked from the user or 0 if it is not selected as liked.
- Content-based recommendations – the content which is liked from the user is used to create user profile, which is later compared to the new contents. Their rank is calculated based on the similarity to the profile. The user profile is created from the keywords and tags contain at objects published, liked or shared by the user. Both user profile and content are represented as N-dimensional keyword vectors extracted from the text content. Vector Space model [12] is used as the weight of the keywords is calculated with TF/IDF metric and similarity between them is calculated with cosine similarity metric.
- Social recommendations – it is based on the social interaction between the user and its friends. The approach differs from collaborative recommendation as it use only the friend's "liked" objects. Comments, likes and shares are used to determine the relationship between the users. Relationship with the users who publish the content and users which comment and liked the content is used to determine how interesting it will be for the user.

7 Conclusions and Future Work

Presented research shows that social network's users spend most of their time looking for interesting information. Results clearly show the necessity of additional application that helps users at their everyday interaction at the social network and order their news feed based on their preferences. Authors already have developed an application which offers filtering options, automated and manual settings on the different sources of content. Next steps are beta testing and spreading at the social network.

Acknowledgments

This work was supported by the European Social Fund through the Human Resource Development Operational Programme under contract BG051PO001-3.3.06-0052 (2012/2014).

References

1. Internet Usage Statistics. http://www.internetworldstats.com/stats.htm (visited June 2013)
2. Kiss, J.: Facebook hits 1 billion users a month. The Guardian, October 2 (2012)
3. One billion – key metrics. http://newsroom.fb.com/download-media/4227 (visited June 2013)
4. Bennett, J., Lanning, S.: The Netflix Prize. KDD Cup and Workshop (2007)
5. Dror, G., Koenigstein, N., Koren, Y., Weimer, M.: The Yahoo! Music Dataset and KDD-Cup'11. urlwww.eng.tau.ac.il/ noamk/papers/DKKW11.pdf (visited August 2013)
6. Antulov-Fantulin, N., Bošnjak, M., Žnidaršić, M., Grčar, M., Morzy, M., Šmuc, T.: ECML-PKDD 2011 Discovery Challenge Overview. In: Proc. of ECML-PKDD 2011 Discovery Challenge Workshop. (2011) 7–20 http://www.matko.info/publications/ecmlpkdd11-challenge-overview.pdf
7. Chechev, M.: Recommender systems challenges. KDD Cup 2011. Days of the Science, V. Turnovo (2011)
8. KDD Cup 2012. http://www.kddcup2012.org/ (visited June 2013)
9. Massa, P., Avesani, P.: Trust-aware collaborative filtering for recommender systems. In: Proceedings of the International Conference on Cooperative Information Systems. CoopIS (2004) 492–508
10. What's the Difference between Top News and Most Recent? http://www.facebook.com/blog/blog.php?post=414305122130 (visited June 2013)
11. Facebook Graph Api. https://developers.facebook.com/docs/reference/apis/ (visited June 2013)
12. Salton, G., Wong, A., Yang, C.: A vector space model for automatic indexing. Communications of the ACM **18**(11) (1975) 613–620
13. Lemire, D., Maclachlan, A.: Slope One Predictors for Online Rating-Based Collaborative Filtering. In: Proceedings of SIAM Data Mining (SDM'05). Newport Beach, California, April 21–23 (2005) http://www.daniel-lemire.com/fr/documents/publications/lemiremaclachlan_sdm05.pdf
14. Cantador, I., Brusilovsky, P., Kuik, T.: 2nd Workshop on Information Heterogeneity and Fusion in Recommender Systems (HetRec 2011), In: Proc. of the 5th ACM Conf. on Recommender Systems (RecSys 2011). Chicago, IL, USA (2011)
15. O'Donovan, J., Smyth, B.: Is trust robust?: An analysis of trust-based recommendation. In: Proceedings of the 11th International Conference on Intelligent User Interfaces IUI'06. ACM, New York, NY, USA (2006) 101–108
16. Fuguo, Z., Shenghua, X.: Analysis of trust-based e-commerce recommender systems under recommendation attacks. In: Proceedings of the The First International Symposium on Data, Privacy, and E-Commerce ISDPE'07. IEEE Computer Society, Washington, DC, USA (2007) 385–390
17. Dunbar, R.: How many friends does one person need? Dunbar's number and other evolutionary quirks. Faber and Faber (2010)
18. Epinions.com: Product Reviews and Consumer Reports. http://epinions.com (visited June 2013)

Increasing the Level of PNG Compression of Grayscale, Natural Images by Applying Noise Reduction Filters on Color Data

Veselina Vetova-Pavlova

Faculty of Mathematics and Informatics, Sofia University “St. Kliment Ohridski”,
5, James Bourchier Blvd., 1164 Sofia, Bulgaria
vvetova@fmi.uni-sofia.bg

Abstract. PNG is a well known lossless format. The compression method implemented in it is Deflate/Inflate and depends strongly on the entropy of the data. Natural images, where a lot of visually redundant information is encoded, have high entropy and low level of compression when PNG is used. To effectively decrease the entropy of natural images and preserve their quality, a noise reduction filters can be used. This research aims to evaluate the entropy reduction effect and the reduction of the PNG files size for a set of 25 noise reduction filters. In addition, it examines the changes in the entropy when different filter parameters like size are changed.

Keywords: PNG compression, natural images, noise-reduction filters, entropy

1 Introduction

PNG is a portable, legally unencumbered, well-compressed, well-specified standard for lossless bitmapped image files [1]. The format is fully streamable with a progressive display option and provides better compression in most of the cases, alpha channels (variable transparency), gamma correction and two-dimensional interlacing [2]. It can be used to compress both color and grayscale images. Currently PNG is the most popular lossless image compression format on the World Wide Web [3].

The compression method implemented in PNG is Deflate/Inflate [4]. The information is first compressed with LZ77 [5] and then with Hufmann coding [6]. The data stream is encoded in ZLIB format [7]. The algorithm is highly reliable, provides good compression, good encoding speed, excellent decoding speed, minimal overhead on incompressible data, and modest, well-defined memory footprints for both encoding and decoding [1]. The level of compression of both algorithms depends strongly on the entropy of the data [6, 5]. Usually signals with lower entropy have better compression.

The entropy is a measure of uncertainty of a random variable [6]. The entropy of an image is calculated as a first order histogram characteristic [8]

$$S_E = -\sum_{b=0}^{L-1} \frac{N(b)}{M} \log_2 \frac{N(b)}{M} \tag{1}$$

where L is the number of shades of gray, b is the number of pixels that have a value b and M is the number of all pixels in the image. Images that have less colors usually have lower entropy. PNG compression includes also a pre-compression stage called filtering [1] where the entropy of the image is reduced further by processing the pixels with one of five different algorithms.

A natural image is an image that is obtained by a sensor system, for example a camera or a scanner. Usually the resulting image has significant amount of visually redundant information. The main reasons is that the neighboring pixels in homogeneous regions of the image have small, often nonuniform, variations that are invisible for a human observer [9]. Thus even images that visually consist of a few colors may contain several hundreds of values with small variance that will increase the entropy.

The size of a natural images can be reduced by using a lossy algorithm. Those algorithms achieve better compression at the price of quality loss. JPEG is the most popular algorithm for lossy image compression and is widely used in the World Wide Web [3]. The main problem of the standard is that blocking artifacts are inserted in the resulting image. These artifacts often are clearly visible in homogenous regions or in regions that are filled with certain pattern.

To reduce the variance between the neighboring pixels, a natural image can be presented as an image that is contaminated with certain noise. Then different noise reduction techniques can be used to reduce the entropy and improve the PNG compression. There is a great variety of noise reduction algorithms [8, 9, 11]. However, two major restrictions have to be fulfilled when choosing appropriate method. First, only methods that effect uniformly the entire image will be effective since they will preserve the uniform and pattern filled areas. Second, the visual differences between the original and the processed image have to be minimal in order to insure sufficient quality of the result.

The performance of a certain noise-reduction filter depends highly on the characteristics of the textures included in the image, as well as of the filter mask that is used. This paper presents the results of an experiment that was conducted in order to observe the effect of a group of different noise-reduction filters upon the image entropy (1). The filters were applied over two datasets of textures with various characteristics. Five noise-reduction filters where included in the experiment – Gauss, Uniform, Median, Minimum and Maximum [8].

The Gauss and Uniform filters are two of the most popular linear filters. The linear noise reduction filters are a simple, fast and flexible method to reduce uniform noise with low, often nonuniform variation [8], which is often the case with natural images. When applied to an image they produce slight blurring that smooths small variances and might result in reducing the entropy. Also when applied to homogenous regions, these filters tend to produce only small, usually undetectable changes.

A major disadvantage of the linear noise-reduction filters is that the contours of the segments are also blurred. In addition high-frequency noise is not reduced efficiently. To handle these cases three statistical noise-reduction filters were added – Median, Minimum and Maximum. Although these filters are slower than the linear filters, the Median filter preserves the contours better than most of the linear filters while the Minimum and Maximum filters can handle high-frequency noise with sufficient quality [8].

For each filter a set of different masks was chosen to observe the influence of the mask changes upon the image entropy. A total of 25 different filters were tested – 8 Gauss, 2 Uniform and 5 of each of the statistical filters. The changes in the entropy and in the size of the PNG file were calculated for each filter and then analyzed in order to provide dependencies that can be used to determining the effectiveness of each of the noise reduction filters.

2 Methodology

During the experiment two linear filters were tested – Uniform and Gauss filters. Only filter masks with size of 3 and 5 px were used in order to reduce the influence of the filter over the visual quality of the image. In addition the influence of the σ parameter of the Gauss filter was observed. Since higher values of σ result in blurrier images, only values that were less or equal to the size of the mask were included. A total of 8 masks for Gauss filter were included in the experiment – with size of 3×3 px and $\sigma = 1, 2, 3$ and with size of 5×5 px and $\sigma = 1, 2, 3, 4, 5$.

Each of the statistical filters included in the experiment was calculated for five different masks, resulting in 15 different filters. To reduce the influence of the filter over the visual quality of the image only four-symmetric masks with size of 3×3 px and 5×5 px, presented in Fig. 1, were used during the experiment. The values of pixels colored in gray were taken in consideration.

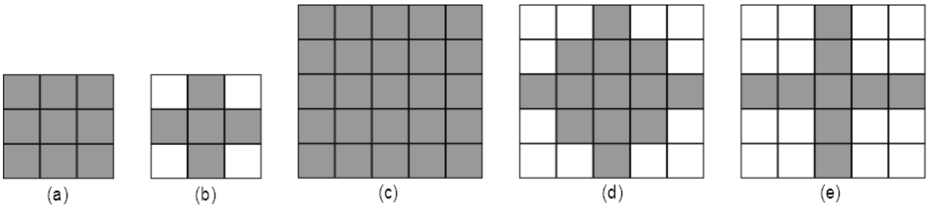


Fig. 1. Masks used to calculate statistical filters: (a) 3×3, Full mask, (b) 3×3, Cross/rhombus mask, (c) 5×5, Full mask, (d) 5×5, Rhombus mask, (e) 5×5, Cross mask

During the experiment two independent datasets of textures with different properties were used. All test images represent natural images of uniformly or close to uniformly distributed textures. Fullcolor images were converted to grayscale in advance. The 2.1D Texture Dataset [13] consists of 80 images of homogenous natural textures with size at least 200×140 px and entropy between 4 and 7.5. Images 2.pgm, 5.pgm, 10.pgm, 39.pgm, 56.pgm, 64.pgm, 65.pgm and 75.pgm were excluded as text was included in them that was not part of the texture and would effect the image histogram and entropy. The VixTex [14] is a texture dataset that includes texture scenes and texture patches. During the experiment only the texture patches with size 512×512 px were used. The entropy of the images varies between 4 and 7.5. Images Buildings.0000.ppm – Buildings.0010.ppm, Tile.0005.ppm and Tile.0006.ppm were excluded as large areas of different textures were included that would significantly change the image histogram and entropy. The results generated by the execution of the experiment on the datasets were analyzed separately in order to be compared.

The experiment consisted of three phases – filtering, sampling and entropy calculation (Fig. 2).

Filtering
Sampling
Entropy Calculation

Fig. 2. Phases of the experiment

In the first phase the images in the datasets were filtered separately with each of the filters. In result for each of them a total of 26 PNG images were created – 25 filtered images and the original. For each of them the size of the resulting file was calculated.

In the second phase for each original image 40 square samples in four different sizes, with random centers, were taken. The maximum sample size was determined as the maximum power of two that is smaller than both half of the width and half of the height of the image. The other three sample sizes were calculated as 1/2, 1/4 and 1/8 of the maximum size. For each size 10 different samples were taken. The selected sizes ensure that non of the samples will be bigger than a quarter of the original image, to ensure variety and each of them will be big enough to contain sufficient number of texels.

In the third phase the entropy (1) of each sample was calculated for the original and the filtered images that were generated in the first phase. The average entropy of each image, original or filtered, was calculated as an average of the entropies of its samples, to ensure translation invariance.

In order to analyze the behavior of the filters, the experimental results for each filter and each dataset of textures were grouped in two sets. The set $S_{\text{filter distribution}}$ is focused on the distribution of the entropy of the filtered images with respect of the entropy of the originals. It includes a two-dimensional element of the type (original entropy, filtered entropy) for each original image in the respective texture dataset. Linear regression analysis was performed and for each filter set the parameters of the approximation line were determined, and the minimum absolute error was calculated.

The set $S_{\text{filter ratio}}$ is focused on the ration between resulting and the original entropy of the images. It includes a fraction $\frac{\text{filtered entropy}}{\text{original entropy}}$ for each image. Then, for each filter the range, the mean absolute deviation, the variance and the standard deviation were calculated [12].

During the experiment Python 2.6 was used as a programming language. All image processing operations except the filtering were performed with the libraries PIL 1.1.7, Scipy 0.12.0b1 and Numpy 1.7.0. The linear regression was performed with the function polyfit of the library Numpy with degree = 1.

3 Results

All filters achieved reduction of the entropy in more than 75% of the images and reduction of the size in more than 79% of the images in both datasets. The size of more than 50% of the tested images was reduced to a half or less of the original size by at least one of the filters. In some cases the size was reduced to 25% of the original. The average percent of entropy reduction for a filter, varies from 1.3% to 13.2% for the images in 2.1D Texture Dataset and from 2.3% to 16.7% for the images in VixTex Texture Dataset. Accordingly the average percent of PNG file size reduction varies from 4% to 45.7% for the images in the first texture dataset and from 4.5% to 50.8% for the images in the second. The percent of images with reduced entropy is high for all filters in both datasets. It varies from 88% to 100% for the images in 2.1D Texture Dataset and from 75% to 100% for the images in VixTex Texture Dataset.

In both datasets, the statistical filters achieved better percent of success in entropy reduction than the linear filters, with only a few exceptions. The tendency also stands for the average percent of file size reduction, although the linear filters perform slightly better. The results for the average percent of entropy reduction are similar. An interesting observation is that the linear filters achieve clearly higher average percent of file size reduction. Although these filters reduce the entropy poorly in comparison with the statistical, they might improve the pre-compression filtering phase and thus to reduce indirectly the entropy of the final data-stream.

With small exceptions, for linear filters the average percents of entropy and file size reduction for both datasets increase when the size of the filter mask increases. For Gauss filters increasing the value of σ has the same effect. Increasing the size or the number of the pixels included in the mask of the statistical filters produces similar results. These tendencies are shown in Table 1.

The elements from the set $S_{\text{filter distribution}}$ for each filter have nearly linear behavior. Table 1 presents also the parameters of the approximation line for each filter. The mean absolute error(MAE) [12] varies from 0.06 to 0.35 for the 2.1D Texture Dataset and from 0.08 to 0.45 for the VisTex Texture Dataset. In general, the statistical filters produce higher MAE. For all filters the mean absolute error and the bias of the approximation line increase with the increasing

Table 1. Percent of entropy and size reduction. Parameters of the approximation lines

	2.1D Texture Dataset					VixTex Dataset				
Filter name	Entr. reduct. %	Size reduct. %	MAE	Line bias	Line slope	Entr. reduct. %	Size reduct. %	MAE	Line bias	Line slope
Gauss 3 × 3, $\sigma = 1$	98.613	82.95	0.093	0.471	0.941	97.668	81.666	0.112	0.808	0.894
Gauss 3 × 3, $\sigma = 2$	98.502	81.838	0.096	0.484	0.94	97.447	80.408	0.12	0.866	0.886
Gauss 3 × 3, $\sigma = 3$	98.456	81.57	0.096	0.487	0.94	97.377	80.08	0.122	0.879	0.885
Gauss 5 × 5, $\sigma = 1$	98.024	77.395	0.107	0.491	0.944	96.645	76.272	0.143	1.035	0.866
Gauss 5 × 5, $\sigma = 2$	97.447	72.096	0.124	0.545	0.941	95.594	70.723	0.174	1.288	0.835
Gauss 5 × 5, $\sigma = 3$	97.24	70.561	0.131	0.576	0.938	95.205	68.972	0.185	1.371	0.824
Gauss 5 × 5, $\sigma = 4$	97.119	69.985	0.135	0.598	0.936	95.03	68.258	0.19	1.408	0.82
Gauss 5 × 5, $\sigma = 5$	97.061	69.717	0.137	0.606	0.935	94.934	67.908	0.193	1.428	0.817
Max. 3 × 3, cross/romb mask	97.046	95.91	0.121	0.328	0.979	97.158	95.127	0.178	0.387	0.968
Max. 3 × 3, full mask	94.945	71.901	0.184	0.734	0.936	94.844	71.322	0.261	0.91	0.906
Max. 5 × 5, cross mask	94.794	87.18	0.197	0.763	0.933	94.446	88.41	0.291	1.254	0.853
Max. 5 × 5, romb mask	93.376	87.777	0.226	1.098	0.892	92.91	86.843	0.332	1.688	0.795
Max. 5 × 5, full mask	89.42	54.53	0.298	2.306	0.725	88.119	51.795	0.452	3.35	0.545
Med. 3 × 3, cross/romb mask	97.222	86.003	0.062	0.425	0.961	96.762	85.53	0.083	0.431	0.963
Med. 3 × 3, full mask	96.013	84.284	0.084	0.616	0.943	95.169	83.14	0.124	0.754	0.926
Med. 5 × 5, cross mask	95.684	81.019	0.086	0.66	0.939	94.575	80.143	0.13	0.817	0.921
Med. 5 × 5, romb mask	95.141	80.566	0.099	0.763	0.928	93.913	79.216	0.149	0.993	0.899
Med. 5 × 5, full mask	93.791	75.39	0.122	1.034	0.896	91.755	73.094	0.188	1.4	0.85
Min. 3 × 3, cross/romb mask	95.385	94.08	0.144	0.646	0.945	94.108	91.649	0.146	0.23	1.025
Min. 3 × 3, full mask	92.87	71.154	0.216	1.312	0.861	91.072	69.117	0.215	0.406	1.03
Min. 5 × 5, cross mask	92.427	84.857	0.24	1.557	0.824	89.734	83.168	0.243	0.588	1.014
Min. 5 × 5, romb mask	90.927	85.449	0.271	1.958	0.771	88.33	81.568	0.275	0.72	1.007
Min. 5 × 5, full mask	86.742	54.263	0.352	3.205	0.588	83.254	49.105	0.366	1.338	0.954
Uniform filter 3 × 3	98.407	81.205	0.099	0.482	0.941	97.272	79.58	0.126	0.913	0.88
Uniform filter 5 × 5	96.841	69.257	0.144	0.648	0.931	94.559	67.117	0.202	1.483	0.812

Table 2. Set characteristics of $S_{\text{filter ratio}}$

	2.1D Texture Dataset				VixTex Dataset			
Filter name	Mean	MAD	SD	Range	Mean	MAD	SD	Range
Gauss 3 × 3, $\sigma = 1$	0.986	0.015	0.027	0.189	0.977	0.021	0.03	0.203
Gauss 3 × 3, $\sigma = 2$	0.985	0.016	0.028	0.19	0.974	0.023	0.032	0.213
Gauss 3 × 3, $\sigma = 3$	0.985	0.016	0.028	0.189	0.974	0.023	0.033	0.212
Gauss 5 × 5, $\sigma = 1$	0.98	0.018	0.029	0.196	0.966	0.027	0.038	0.243
Gauss 5 × 5, $\sigma = 2$	0.974	0.02	0.031	0.208	0.956	0.034	0.046	0.28
Gauss 5 × 5, $\sigma = 3$	0.972	0.021	0.032	0.212	0.952	0.036	0.049	0.292
Gauss 5 × 5, $\sigma = 4$	0.971	0.022	0.033	0.214	0.95	0.037	0.05	0.297
Gauss 5 × 5, $\sigma = 5$	0.971	0.022	0.033	0.214	0.949	0.037	0.051	0.298
Max. 3 × 3, cross/romb mask	0.97	0.018	0.024	0.121	0.972	0.028	0.036	0.182
Max. 3 × 3, full mask	0.949	0.029	0.037	0.192	0.948	0.04	0.054	0.319
Max. 5 × 5, cross mask	0.948	0.031	0.04	0.218	0.944	0.046	0.062	0.356
Max. 5 × 5, romb mask	0.934	0.036	0.047	0.251	0.929	0.051	0.071	0.439
Max. 5 × 5, full mask	0.894	0.052	0.067	0.356	0.881	0.069	0.099	0.658
Med. 3 × 3, cross/romb mask	0.972	0.011	0.014	0.075	0.968	0.014	0.021	0.128
Med. 3 × 3, full mask	0.96	0.015	0.019	0.091	0.952	0.022	0.032	0.196
Med. 5 × 5, cross mask	0.957	0.016	0.02	0.094	0.946	0.024	0.033	0.204
Med. 5 × 5, romb mask	0.951	0.018	0.023	0.104	0.939	0.027	0.039	0.235
Med. 5 × 5, full mask	0.938	0.023	0.029	0.125	0.918	0.036	0.05	0.285
Min. 3 × 3, cross/romb mask	0.954	0.023	0.029	0.163	0.941	0.021	0.028	0.156
Min. 3 × 3, full mask	0.929	0.034	0.044	0.236	0.911	0.03	0.039	0.21
Min. 5 × 5, cross mask	0.924	0.037	0.049	0.265	0.897	0.034	0.043	0.211
Min. 5 × 5, romb mask	0.909	0.042	0.056	0.299	0.883	0.037	0.048	0.245
Min. 5 × 5, full mask	0.867	0.058	0.076	0.398	0.833	0.048	0.061	0.305
Uniform 3 × 3	0.984	0.016	0.028	0.19	0.973	0.024	0.034	0.22
Uniform 5 × 5	0.968	0.023	0.035	0.219	0.946	0.039	0.052	0.305

of the mask size while the slope of the approximation line, with small exceptions, decreases. Also increasing the σ value for Gauss filters or the number of pixels included in the mask for statistical filters produces the same effect. The higher MAE might be a result of a higher sensitivity of the respective filters to different characteristics of the images that can be examined in future in order to produce a set of measures that will help choosing the best noise reduction filter for certain texture. Filters with lower line-slope might be most effective for textures with high entropy while filters with higher line-slope but lower bias might proof to be the better choice for textures with low entropy.

These results are further confirmed by the $S_{\text{filter ratio}}$ sets. Increasing the size of the filter masks, the value of σ for Gauss filters or the number of pixels included in the mask for statistical filters decreases the mean value of the respective set and increases its mean absolute deviation (MAD), standard deviation (SD), range and variance (Table 2). The decreasing mean value shows that these filters are more effective in general but the bigger range, variance and deviation are result of the higher sensitivity of the respective filters to the characteristics of the images.

4 Conclusion

The tested noise-reduction filters can be used for reduction of the entropy of certain images and improvement of the PNG compression. The efficiency of the filter depends strongly on the its parameters. The statistical filters reduce the entropy more effectively but the linear filters produce better PNG compression since they improve the performance of the pre-compression filtering stage, reducing indirectly the entropy of the final datastream.

Filters with bigger size, higher value of σ for Gauss filters or the higher number of pixels included in the mask for statistical filters, produce both better average entropy and file size reduction but they also tend to be more sensitive to the characteristics of the respective image.

Acknowledgments.

This work was supported by the European Social Fund through the Human Resource Development Operational Programme under contract BG051PO001-3.3.06-0052 (2012/2014).

References

1. Boutell, T.: PNG (Portable Network Graphics) Specification Version 1.0 (1997)
2. Roelofs, G., Koman, R.: PNG: The Definitive Guide. O'Reilly & Associates Inc. (1999)
3. Gelbmann, M.: The PNG image file format is now more popular than GIF. W3Techs (2013) `http://w3techs.com/blog/entry/the_png_image_file_format_is_now_more_popular_than_gif`

4. Deutsch, L.P.: DEFLATE compressed data format specification version 1.3 (1996)
5. Lin, S.K.: Encyclopedia of Algorithms. Springer-Verlag (2008)
6. Cover, T.M., Thomas, J.A.: Elements of information theory. Wiley-Interscience (2012)
7. Deutsch, J., Gailly, L.: RFC 1950-ZLIB Compressed Data Format Specification version 3.3 IETF/IESG (1996)
8. Pratt, W.K.: Digital image processing: PIKS inside. Wiley (2007)
9. Acharya, T., Ajoy, K.R.: Image processing: principles and applications. Wiley-Interscience (2005)
10. Miano, J.: Compressed Image File Formats: Jpeg, Png, Gif, Xbm, Bmp. Addison-Wesley Professional (1999)
11. Russ, J.C.: The image processing handbook. CRC press (2006)
12. Ghosh, A.K.: Introduction to measurements and instrumentation. Prentice-Hall of India Private Limited (2007)
13. 2.1D Texture Dataset `http://web.engr.oregonstate.edu/sinisa/TexturesDataset.html`
14. VixTex Texture Dataset `http://vismod.media.mit.edu/vismod/imagery/VisionTexture/distribution.html`

Read Optimization Based on Column-Oriented DBMS

Hristo Kyurkchiev and Kalinka Kaloyanova

Faculty of Mathematics and Informatics, Sofia University "St. Kliment Ohridski",
5, James Bourchier Blvd., 1164 Sofia, Bulgaria
{hkyurkchiev,kkaloyanova}@fmi.uni-sofia.bg

Abstract. Most transactional database systems have been optimized for write performance. However, as analyzing and mining transactional data straight from the DBMS has become increasingly popular, read optimization has gathered significant interest recently. In this paper we compare the widely used row oriented DBMS with the column stores, which gain more and more momentum in the last several years. We focus on the data model of both systems as well as specific read optimization techniques. By analyzing previous performance studies we draw conclusions about the advantages and disadvantages of both systems in general and on analytical query loads.

Keywords: database systems, data warehouses, column stores, performance evaluation, analytical query load

1 Introduction

The relational database model introduced by Codd [5] dominates the database scene for the last couple of decades due to its atomicity, consistency, isolation, and durability properties [9]. Most traditional database management systems (DBMS) use the relational model not only as a logical data model, but also as a physical one, storing the data tuples contiguously on the disk [14]. This is usually denoted as N-ary storage model (NSM). Column-stores, on the other hand, although also using the relational model as a logical data model, use the decomposed storage model (DSM) for physical data storage [1]. We analyze the differences between the two approaches and compare their performance on read loads, which are typical for data warehouse systems. On this basis we draw conclusions about their suitability for different setups.

2 Theoretical Background

Both traditional relational DBMS (row stores) and column stores use as a logical model the relational data model [5]. Although well known, for completeness we list its basic components bellow.

2.1 Logical Data Model

The main concepts of the relational model as outlined in [5, 6, 9] are:

- Relation – a subset of the Cartesian product of some sets $D_1, D_2, \ldots, D_n$, usually visualized as a two-dimensional table.
- Attributes – the names of the columns in the table, e.g. $A_1, A_2, \ldots, A_n$.
- Relation Schema – the name of the relation together with the names of its attributes, e.g. $R(A_1, A_2, \ldots, A_n)$.
- Tuple – a value row in the table, e.g. $(a_1, a_2, \ldots, a_n)$ with $a_i \in D_i$.
- Domain – the data type of an attribute (one of the above D_i).
- Key of relation – is a constraint on a given relation that denotes that no two tuples are to agree on (have the same values for) all attributes in the key.

2.2 Physical Data Model

The physical organization of the data of most traditional relational DBMS is based on the NSM [1, 12, 14]. Resulting in essentially storing the data using its logical representation and translating it to physical structures. Thus in traditional relational databases the relations are stored tuple by tuple, with each tuple being a single record on disk and being stored close to the other tuples of the relation [9, 12].

For example, if there is a relation $R_i(S_i, A_1, A_2, \ldots, A_m)$ then with NSM all of the attributes would be stored together, clustered or hashed on the surrogate attribute (e.g. an auto-increment key) [12]. An example of such physical representation of a relation then would look like Fig. 1, where the header is some system information and a single record represents each tuple with all attributes' values in it.

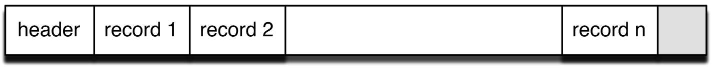


Fig. 1. NSM – a typical block holding records [9]

Column stores use a modification of the DSM [1]. It is a fully transposed model, which also includes surrogates. DSM pairs each attribute value with the surrogate of its conceptual schema record in a binary relation. Each of these binary relations is present twice – sorted or hashed with an index on each of the two attributes [12].

The same relation $R_i(S_i, A_1, A_2, \ldots, A_m)$ in the DSM is decomposed into binary relations, e.g. $R_{i1}(S_i, A_1), R_{i2}(S_i, A_2), \ldots, R_{im}(S_i, A_m)$. If there is another relation R_j, which can be joined to R_i then there is also the binary relation $R_{ij}(S_i, S_j)$, which represents the join and is called "*join index*" [12]. Thus the physical representation of the relation in DSM would look like Fig. 2, where the

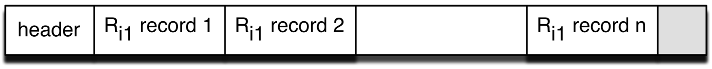


Fig. 2. DSM – physical representation

header is some system information and a single record represents a tuple in some of the binary relations.

From these specifications one can discern that in the NSM there is the need for a single read or write for a single tuple access. This leads to both operations, e.g. whole tuple access and insert or delete statements' execution being regarded as easy and almost equally inexpensive disk operations. For queries such as the ones usually performed in data warehouses where only a few attributes are being aggregated to supply the result, however, the NSM requires reading all tuples and extracting the necessary attributes, which can prove expensive. The DSM on the other hand stores the attributes by themselves (with the exception of the surrogate), making it better suited for aggregated queries, as it would require only to read the records for the binary relations of the needed attributes. It is to be expected, however, that with it the insert and delete operations would be much more expensive as they would require read or write of potentially up to $2m$ blocks to perform (if m is the number of blocks originally). These analytical observations are confirmed by the results of Copeland and Khoshafian [7, 12]. It is also important to note that while the insert and delete performance are worse compared to NSM and the update is relatively on par, the read performance advantages of the distributed storage model is dependent largely on the architecture and on the number of attributes in the retrieval query [7, 12].

2.3 Limitations of DSM

As briefly outlined there are some limitations of the DSM. We focus on the ones connected with read queries. There of course are other limitations, some of which include difficulties when inserting or updating records but they are orthogonal to the read limitations and are a topic of a separate discussion. The limitations that concern read query performance include:

- Join limitations – as the relations are stored in a decomposed state it is expensive to join all of their parts back together when the relation is needed in its original state. Moreover, in typical analytical queries usually one facts table is joined with one or more dimension tables in order to provide the desired result, which again can be expensive when the relations are saved decomposed.
- Read performance limitations – as discovered in the 1980s [7, 12], there is a correlation between the number of selected tuples, the number of accessed attributes and the performance of the DSM. When the selectivity is low (approx. 1%) the number of accessed attributes should also be low (at most 30%) for the performance to be on par with the one for NSM. Almost the

same principle applies to joins when using the DSM. In general the model outperforms the NSM when the number of accessed attributes is below 50% of all attributes in the relation, and/or the number of selected tuples is above 10% of all tuples in the relation.

- Storage limitations – as there is redundancy in the DSM's specifications it requires more storage space than the classical NSM.

Column stores overcome most of these limitations by employing a slightly different model and using some optimizations especially for read queries.

3 Column-Oriented DBMS

Based on the ideas of the DSM the column store research started in the 2000s. It elaborated on the concepts of the decomposed storage model and addressed some of its limitations, while also introducing certain optimizations.

3.1 Data Model

We use C-Store as a basis for the model employed by column stores as most of its concepts are also applied by other implementations (e.g. MonetDB, Vertica).

C-Store decomposes the relations into *projections*, with each projection being *anchored* to a logical relation $\mathrm{R_i}$ and containing one or more attributes from it and possibly one or more attributes from relations $\mathrm{R_j}$, which are in a one-to-many relationships with the relation $\mathrm{R_i}$ (e.g. have foreign key constraints). The tuples are being stored column-wise, e.g. if there are m attributes in a given projection there would be m data structures storing each of them, with all column values being sorted in a left to right order by the *sorting key* – a set of the attributes in the projection [14]. The physical model is similar to the DSM, with the exception that there are no duplicates of each column and there are projections on the logical level, which enable some optimizations. In the data model of C-Store there is also the ability to partition the projections. This is done using segments with each segment being associated to its own unique *segment identifiers*, which are positive integer numbers. Since C-Store supports only value-based partitioning on the sorting key each segment has a key range associated with it [14].

Naturally, to answer a query there should be a set of projections, which cover all of the queried attributes. To be able to join these projections so that the necessary tuples are constructed C-Store uses *storage keys* and *join indices*.

The storage keys represent the position of the column's value in the relation (e.g. the tuple it is in). In the main unit of the C-Store, the read-optimized store, they are not physically stored but inferred, depending on the encoding schema (see below). In the writable store they are physically stored and are represented as integers, higher than the biggest storage key in the read-optimized store [14].

Join indices are used by C-Store as means of tuple reconstruction. Their semantics is that if there are two projections P_1 and P_2 that cover all attributes of a given table T then a join index from the M segments of P_1 to the N segments of P_2 is logically a set of M ordered pairs, one for each segment S of P_1, e.g. (s, k), where s is a segment identifier of a segment in P_2 and k is the storage key in the segment s. This is always an one to one mapping as all projections are anchored to the same table T [14].

3.2 Read Optimizations

Column stores also employ some optimizations that try to address the limitations that we observed in the DSM. While some of them (e.g. projections) are specific to C-Store, most are common among column-oriented DBMS. These include:

- separate read-optimized and writable stores;
- projections;
- compression and data encoding;
- block iteration;
- late materialization strategies;
- invisible joins and join indices.

Separating the read-optimized and writable stores into two distinct pieces of software alleviates the problem of optimizing the structure for read mostly queries, while at the same time providing adequate insert and update services [14]. Although this is mostly a write related optimization, which as we already mentioned is not the focus of this paper it also has some effect on read performance. Mainly, when there is no need to support the full spectrum of inserts and updates the store can employ more aggressive encoding schemas, more densely packing the values, etc.

Projections can be seen as an optimization when C-Store is compared with DSM. The basis for this is that although they can provide redundancy (if one attribute is present in two projections) the model does not enforce it explicitly, thus minimizing disk space usage. Moreover projections can be build in such a way that they provide the necessary attributes for often executed queries, which lifts the need to perform expensive joins and tuple reconstructions based on the whole relation.

Compression provides advantages such as low disk usage. Although this is not so significant nowadays as the cost of hard drives drops continuously, it also has the effect of minimizing I/O lag and thus speeding up queries [4]. The compression/encoding algorithm to be used highly depends on the data. Usually algorithms work best on data with low information entropy. Thus the sorted (and secondary sorted) order of the projections in C-Store is very suitable for employing compression on the data. Prior research [2] has shown that there is no one best algorithm for all data in order not to impair read performance, however, heavyweight compression should be avoided, at it provides smaller size data, but at a higher CPU decompression cost at query execution time. C-Store, itself employs four encoding schemas depending on the data ordering and number of distinct values, which are presented in Table 1.

Table 1. Encoding schemas used in C-Store [14]

Distinct values/Order	Self-order	Foreign-order
Few distinct values	Columns are represented by an ordered triple (v, f, n), where v is the value, f is the position in the column where v first appears and n is the number of times v appears.	Columns are represented by an ordered pair (v, b), where v is the value and b is a bitmap indicating the positions in which the value is stored. Each bitmap is then run-length encoded to save space.
Many distinct values	Columns are represented with values, which are essentially the deltas from the previous value.	In this case the data is left as it is.

Block iteration is common in most column stores. The idea essentially is that column values are sent to function operators in blocks thus minimizing tuple overhead, which is common in row-oriented DBMS due to their working on function operators on a tuple-per-tuple basis. This approach also provides opportunities for parallelizing the query execution [4].

Late materialization is a strategy for reconstructing tuples. Most times during the query's execution the tuples must be reconstructed in order for the results to be calculated. Naïve column-oriented DBMS use early materialization when executing queries – they join together the columns relevant for the queries early on and then perform the normal row-store operators on the reconstructed tuples. On the contrary in C-Store late materialization is employed, leading to operators being performed directly on the columns themselves and the tuples being reconstructed much later in the query's execution plan from the already computed results. In order to deliver even better performance the positions of the columns, which satisfy the selectivity conditions are saved so that only those values are passed on [4]. This is beneficial in several ways. When using aggregations, which are typical for analytical queries, as well as when there are some selectivity operators, the construction of some tuples is unnecessary. With late materialization these tuples would be skipped, while with early materialization, they would be first materialized and then filtered, which presents overhead. Late materialization is able to operate on the compressed data without the need to first decompress it and reconstruct the tuples. Another benefit is the cache performance, which is due to better cache utilization, as the cache is not polluted with unnecessary attributes' values. Lastly the block iteration described earlier works better on fixed length attributes [4]. There are cases, however, when early materialization is better as prior research [3] suggests. These include high selectivity, non-aggregated queries. Usually though, analytical queries do not fall in these categories as they most often include aggregated data for specific periods of time.

Invisible joins are essentially late materialization for joined queries. Most analytical data warehouse queries include joins between a facts table and some dimensions tables with some aggregation. The idea of invisible joins is that the join conditions are rewritten to appear as selecting conditions of the facts table. The query is then executed on the facts table and when all predicates and operations are performed are the dimension table's tuples reconstructed and joined to those from the facts table [4]. The **join indices** themselves can also be looked at as optimizations as their purpose is to be able to quickly reconstruct tuples so that better read queries performance is delivered.

The model and optimizations described here address the limitations we extracted above. The join limitations are addressed by the invisible joins and join indices. Since projections can include attributes from more than one relation they too can be seen as optimizations, which address join limitations in the cases when the joining relations and their attributes are a part of a single projection. All of the enlisted optimizations concern read optimized performance; even compression can be included in this category since it results in less I/O operations, which are the slowest part in a query's execution. The fact that the model does not explicitly force duplication of the projections and the usage of compression and encoding schemas make it require less storage not only compared to the DSM but also to NSM sometimes by an order of magnitude [2].

4 Read Queries Performance Comparison

We should note that the original experiments [7, 12] have been carried on with different schemas for the ones used for the column stores vs. row stores comparison [10, 11, 14]. More importantly the gap in years implies differences in technology capabilities and restrictions. Since there are no more recent examples for DSM, to the best of our knowledge, we would use the original papers for reference. A comparison with comparable hardware of both approaches is also of academic interest but is a topic for another paper.

4.1 General Read Performance

As already mentioned the results for DSM [7, 12] show that using relation decomposition improves performance only when the selectivity is relatively low (e.g. between 1% and 10%). Even then there are restrictions to the projected attributes (e.g. between 30% and 70%) depending on the actual selectivity, in order for the model to outperform the NSM.

Column stores are an evolved version of the DSM with a different model and optimizations that we already discussed. The performance comparisons between them and row stores also show that the highest advantage is present when there is selectivity only of a fraction of the attributes of the relations in a given query [10, 11, 14]. The proportional relation between selectivity and projected attributes also stands true. There are, however, some improvements as for example for 10% selectivity one study [11] reports a 85% tuple length as a crossover point for the better performance of column stores vs. row stores. This is significantly higher than the 70% of DSM.

4.2 Analytical Queries Performance

For typical data warehouse queries that use only several of a relations attributes [8] (e.g. less than 10%) such as the ones proposed in the Star Schema Benchmark [13] both the DSM and C-Store deliver better performance as compared to traditional relational DBMS. The DSM reaches 25 order of magnitude better performance than the NSM when selectivity is 10% and the number of projected attributes is less than 10% [12]. At the same time C-Store as an implementation of a column store is on average 164 times faster than a row store using NSM on similar conditions as the DSM (e.g. no materialized views employed by the row store).

On the scenario, when there are materialized views in the row oriented database, C-Store delivers an order of 3 magnitude better performance as compared to row stores [4]. The benefits were due to the late materialization optimization, which improved the performance by a magnitude of 3 and the compression, which improved the performance by a magnitude of 2. For some cases [14] this gap is even larger, reaching ten fold performance benefits for specific queries with only a fraction of the size of the database ($\sim$ 2 GB vs. $\sim$ 12 GB).

These results clearly show the superiority of column stores over row stores and even on databases using the DSM, when analytical queries are concerned.

5 Conclusion

In this paper we compared the storage models of row and column stores. Based on this overview we extracted the limitations of the DSM model with regards to read intensive environments. The data model together with the optimizations proposed by the creators of C-Store, however, has overcome most of these as we showed in the analysis. The limitations that remain unresolved include the slower performance as opposed to row stores when the number of queried attributes is high or the selectivity is too high and there are no aggregations. These concerns are not relevant, however, when it comes to analytical query loads. As the read performance comparison has indicated, column stores (in the face of C-Store) outperform traditional row oriented databases by an order of magnitude, when the focus is on analytical queries. This makes them the better choice when selecting a DBMS for read mostly environments.

Acknowledgments

This paper is supported by Scientific Research Fund of the Sofia University "St. Kliment Ohridski" under contract No. 43/2013 and European Social Fund through the Human Resource Development Operational Programme under contract BG051PO001-3.3.06-0052 (2012/2014).

References

1. Abadi, D. et al.: Column-oriented database systems. In: Proceedings of the VLDB Endowment **2**(2) (2009) 1664–1665
2. Abadi, D. et al.: Integrating compression and execution in column-oriented database systems. In: Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data SIGMOD'06 (2006) 671
3. Abadi, D.J. et al.: Materialization Strategies in a Column-Oriented DBMS. In: Proceedings of ICDE 2007, Istanbul, Turkey (2007) 466–475
4. Abadi, D.J., Madden, S.R.: Column-Stores vs. Row-Stores?: How Different Are They Really? SIGMOD (2008) 967–980
5. Codd, E.F.: A relational model of data for large shared data banks. Communications of the ACM **13**(6) (1970) 377–387
6. Codd, E.F.: The Relational Model for Database Management: Version 2. Addison-Wesley (1990)
7. Copeland, G.P., Khoshafian, S.N.: A Decomposed Storage Model. In; Proceedings of the 1985 ACM SIGMOD International Conference on Management of Data **14**(4) (1985) 268–279
8. Dewitt, B.D.J. et al.: How to Build a High-Performance Data Warehouse. (2010) `db.lcs.mit.edu/madden/high_perf.pdf`
9. Garcia-Molina, H. et al.: Database Systems: The Complete Book. Pearson Prentice Hall, Upper Saddle River, New Jersey 07458 (2009)
10. Halverson, A. et al.: A Comparison of C-Store and Row-Store in a Common Framework. In: Proceedings of the 32nd VLDB Conference, Seoul, Korea (2006) Paper Id: 335 `pages.cs.wisc.edu/~alanh/tr.pdf`
11. Harizopoulos, S. et al.: Performance Tradeoffs in Read-Optimized Databases. In: Proceedings of the VLDB Endowment (2006) 487–498
12. Khoshafian, S. et al.: A Query Processing Strategy for the Decomposed Storage Model. In: Proceedings of the Third International Conference on Data Engineering (1987) 636–643
13. Neil, P.O. et al.: The Star Schema Benchmark (SSB) January, 1–10 (2007)
14. Stonebraker, M. et al.: C-store: a column-oriented DBMS. In: Proceedings of the 31st VLDB Conference (2005) 553–564

Denoising of Metagenomic Data from High-Throughput Sequencing

Milko Krachunov

Faculty of Mathematics and Informatics, University of Sofia “St. Kliment Ohridski”,
5, James Bourchier Blvd., 1164 Sofia, Bulgaria
milkok@fmi.uni-sofia.bg

Abstract. Metagenomics is a growing research field focused on the study of genomic data collected from heterogeneous microbial environments. Experimental results are highly sensitive to errors in the data, and the commonly used error filtering approaches are crude and often rely on throwing away a lot of legitimate data to remove noisy reads. This paper presents an original analytical approach to denoising of metagenomic data which attempts to improve the quality by filtering out spurious corrections that would be made by frequency-based error detection, together with an on-going effort to implement an alternative approach using neural networks. Indirect approaches to validate the instances of detection and correction are proposed and tested. The proposed approaches are being developed inside a metagenomic data processing workflow library which will be released as a free software package.

Keywords: metagenomics, denoising, NGS data analysis workflow, artificial neural networks

1 Introduction

1.1 Metagenomics

Metagenomics studies the microbial genetic data found in mixed samples collected from heterogeneous biological environments, such as soil, water basins, and the insides of macro-organisms. The communities of microorganisms in these environments are still largely unexplored, and even the diversity of the species is unknown and the primary subject of much of the research. The in silico experiments pose many unsolved technical and methodological challenges.

From scientific perspective, the comparative analysis of microbial communities is critical for many studies in biology and medicine, in issues ranging from human health [1] in the study of impact of human-borne microbes, to the bacterial and viral evolution [2] in antibiotic use or as a narrow view at the evolution of the species in general.

From the perspective of Informatics, metagenomics is focused on the processing of sequences (strings) of bases from a four-letter alphabet (A, C, G, T), whose content is variable but similar due to the evolutionary relationships between the studied microbes. The sequences encode biological information in a manner that is largely unstructured and chaotically organised, which makes it more difficult to formalise beyond a certain level.

Metagenomics researches are forced to deal with a variety of challenges [3]. A lot of data processing and research projects suffer from the lack well-established approaches to use, which can be further exacerbated by the large datasets produced by high-throughput sequencing equipment. Obstacles are encountered in sequence alignment, which is crucial for performing any form of analysis on the data; error detection and correction, which can lead to significant changes in experimental results; and the choice and combination of software packages to perform processing and analysis.

1.2 Goals

The main goal of the presented project is the development of an improved error detection approach that is more reliable, and more suitable to be extended to perform error correction.

An original analytical error detection and correction approach has been presented. It is based on counting the per-column frequencies of occurrences, using the pair-wise local similarities between the sequences as weights to account for the heterogeneous nature of the datasets, and filter out any spurious error flaggings of rare sequences. Due to the difficulty in obtaining test data, the validation of the error detection has been performed using an indirect validation approach that relies on simulated errors.

On-going work is focused on the addition of an alternative approach that is based on artificial neural networks, presented in this paper, in attempt to find a more precise indicator for the errors that is closely tuned to the characteristics of the data, at the expense of less clarity in the process by which errors are identified.

As a secondary goal of the project, a metagenomic workflow execution library has been developed, which manages the execution of sequence clustering and sequence alignment software packages, performs error detection, correction and validation with the proposed methods and is being reworked as a more generalised software library to handle a wider variety of tasks [4].

Experiments have been carried out to show that the results of filtering of spurious corrections using local similarity weights are consistent with an actual improvement in the quality of detection, making it a viable improvement in denoising attempts.

2 Material and Methods

2.1 The Input Data

16S RNA is often used for metagenomic analysis. It is highly conserved between microbial species, making it useful to do cross-species comparisons; at the same it contains hypervariable regions that can be used to highlight any differences between the sequences, in particular to identify the species, strains, and individuals or map their evolutionary relationships. This makes suitable for studies of the biodiversity as well as studies of phylogenetic trees [5].

Our sample datasets contain short reads between 300 and 500 bases in length, divided in sets of tens of thousands of sequences – between 20000 and 50000 after filtering them by length (≥ 300, ≤ 500 bp) and quality (by throwing out ambiguous bases). All our sample datasets were sequenced using the 454 platform by Roche. It is very suitable for metagenomic experiments because it produces short reads of sufficient length.

2.2 Data Preparation

Before any meaningful column-wise analysis can be performed on the data in a metagenomic sample, the sequences need to be prepared, which includes performing multiple sequence alignment and often sequence clustering [6]. Displacement due to biological or erroneous insertions and deletions make it impossible to directly compare sequences in each column before the sequences are aligned. In addition, due to too much variety in the samples, clustering is often required for getting meaningful results.

The project mainly relies on external tools for the alignment, but due to deficiencies in these tools, a makeshift aligner combining a couple of packages in included in the software and used in our processing. The aligner clusters individual sequences with CD-HIT-454 [7], aligns individual clusters with MAFFT [8] or MUSCLE [9] and then merges them through a custom procedure [10, 4]. The need to resort to such makeshift solutions during metagenomic processing and analysis is the main motivation for the on-going work to rework the project software into a flexible workflow library that can be used for solving problems outside the scope of the project.

2.3 Analytical Approach to Improve Error Detection and Correction

Removing the noise is a significant problem in metagenomic studies. Sequencing equipment produces a large amount of errors that cannot be easily identified, and ignoring them can lead to large deviation in any experimental results. An additional problem is that errors can often be mixed up with legitimate variations between the sequences – such as mutations and the occurrence of rare subsequences. Such variations are important subject in metagenomic studies, and filtering them out as errors is counter-productive.

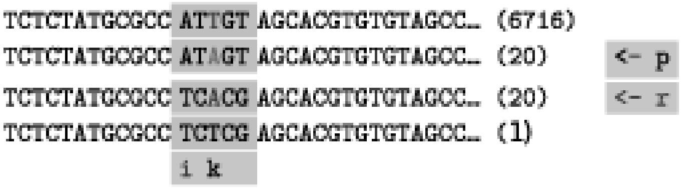
Our analytical approach for error detection, which has been discussed in even further detail in [10, 4], is outlined below.

The Naïve Approach. A common approach to identify correct sequences is to count the per-column rates of occurrence of the four bases. If a base is rare in a column, it is a potential error. For example, when consensus sequences are constructed from homogeneous datasets in de novo sequences, only the base with the highest frequency of appearance is considered correct [11]. After introducing an error threshold to compare against the frequency of occurrence, we denote this the naïve approach which will serve as a “base” for the “improved” approach.

To mathematically express this, we define a score function (representing the base frequency) that we will compare against the error threshold. If we denote the set of reads with R and their count with $n = |R|$, the “naïve” score for position k in read r will be:

$$\operatorname{score}(r,k) = \sum_{p \in R}^{p \neq r} \frac{[r_k = p_k]}{n-1} = \frac{\displaystyle\sum_{p \in R}^{p \neq r} [r_k = p_k]}{\displaystyle\sum_{p \in R}^{p \neq r} 1}. \tag{1}$$

The so-defined score(r, k) yields the portion of reads (excluding the read r) that contain the base r_k at column k. This will count all 39 matching reads for the example in Fig. 1.



```
TCTCTATGCGCC ATTGT AGCACGTGTGTAGCC… (6716)
TCTCTATGCGCC ATAGT AGCACGTGTGTAGCC… (20)      <- p
TCTCTATGCGCC TCACG AGCACGTGTGTAGCC… (20)      <- r
TCTCTATGCGCC TCTCG AGCACGTGTGTAGCC… (1)
             i k
```

Fig. 1. Similarity-weighted error detection

Extending this to error correction is straight-forward – any base with naïve score below the threshold is replaced with the base out of A, G, T and C with the highest score at the same position.

The Similarity-Weighted Approach. To account for the competing distinct sequences found in the dataset, we propose a method to filter out any spurious corrections performed by the naïve approach by introducing the local pair-wise similarity as a weight in (1). This improvement will guarantee that the context around the evaluated bases will be taken into account when making the decision about detection or correction. Dissimilar sequences will be evaluated as different groups, while avoiding corrections to chains of mismatches that are more likely to be rare sequences.

The similarity-weighted approach creates a window around each evaluated base, for each pair of compared sequences estimates the local similarity in the window, and when counting a match or mismatch between the sequences, this local similarity is used as a weight.

To express this mathematically, we amend the formula from (1) to get the "similarity" score function:

$$\operatorname{score}(r,k) = \frac{\displaystyle\sum_{p\in R}^{p\neq r} \operatorname{similarity}(r,p,k)[r_k = p_k]}{\displaystyle\sum_{p\in R}^{p\neq r} \operatorname{similarity}(r,p,k)}. \tag{2}$$

An exponentially decreasing function is used for the similarity. If q is an experimentally evaluated decay parameter, and w is the size of the window, we will use the following function:

$$\operatorname{similarity}(r,p,k) = \frac{\displaystyle\sum_{i\in\operatorname{window}(r,k)} q^{|m-i|}[r_i = p_i]}{\sum_{i\in\operatorname{window}(r,k)} q^{|m-i|}} \tag{3}$$

$$\operatorname{window}(r,k) = (\{k-w,\ldots,k-1,k+1,\ldots,k+w\}) \cap \{i : \exists r_i\}. \tag{4}$$

The window size w is chosen to ensure reasonable execution time (e.g. $w = 10$), as the decay parameter q significantly decreases the effect of the bases near the edges of the window. Thus specified, the score(r,k) from (2) counts only the 19 matches in similar reads in Fig. 1, unlike (1) which counted all 39.

Extending this to correction can be done the same way as for the naïve approach – bases with scores below the threshold can be replaced with bases with the base that would yield the highest score in the same position.

2.4 Supplementing the Analytical with a Neural Network Approach

The inference potential of artificial neural networks can also be utilised to evaluate errors, because it would be able to find more characteristics in the data that are chaotic in nature and much more difficult to formalise or reasonably assess. We are in the process of developing a supplemental neural network that, given a sequence and a position, assesses whether the position is correct or not based on an input describing the evaluated sequence together with the rest of the sequences in the dataset.

Data Input. The biggest challenge in designing a neural network to tackle the problem with error detection is selecting how to provide the input data to the network. It needs to be summarised by criteria that would appropriately represent the information that is most significant for identifying errors, and separated in way that would preserve the meaningful relationships.

Input Generalised by Similarity and Offset. The input is split into bins using multiple similarity classifiers. Each splits the sequences into different sets of bins. A window is created around the evaluated position, and a different similarity is estimated at each position in the window. For each classifier, three inputs neurons (corresponding to 0%, 50% and 100% similarity at the given offset) receive the portion of sequences with a match for the evaluated position as input, as it is shown in Fig. 2.

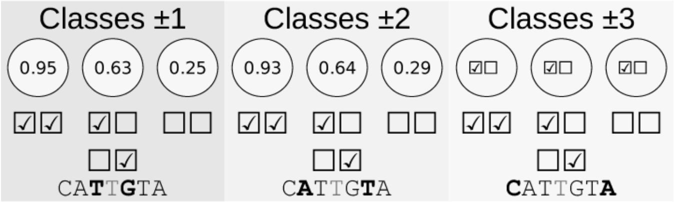


Fig. 2. Example neural network input

The main flaw of this generalisation is that the neural network will not have visibility of the relations between adjacent positions, which after some experimentation proved to be significant in certain cases, such as repeats where the sequencing equipment tends to over- or underestimate the times a single base is repeated in a sequence, and do so in a very specific pattern.

One of the remaining tasks in the project include finding a way to eliminate this disadvantage, e.g. by introducing input neurons providing information on the neighbours of the evaluated position, or input neurons for the similarity in pairs, triplets and so on, but these early suggestions would significantly complicate the neural network design.

Cost Function. Due to the lack of good representative training and test data, the neural network is trained against the simulated errors produced by the indirect validation approach discussed below. If p is the neural network output, the cost function will be:

$$E(p) = \begin{cases} (1-p)^2, & \text{simulated error,} \\ (p)^2, & \text{no simulated error.} \end{cases} \tag{5}$$

A more complex function taking into account the mismatch between the global error rate and the errors detected by the network can be formulated to more closely match the validation method, but this could introduce additional unwanted biases.

Topology. The initial topology used for the first experiments is a feed-forward network with a single hidden layer and error back-propagation with mean squared error. The input layer is represented with $3 \cdot w$ neurons (w being the window radius, and 10 being the initial choice for w), $4 \cdot w$ hidden neurons and a single output neuron. Until the input and cost function are settled, experiment will be performed varying the rest of the parameters of the network.

2.5 Validating Error Detection and Correction

The lack of a large number of reliable test and training dataset created the need to utilise indirect approaches to validate the analytical method and train the neural network. Two proposed such approaches are described in [4], and the details of one of them are briefly discussed here.

Validation through Repeated Application Approach. The main validation approach to compare two ("base" and "improved") correction methods performs error correction on the raw data, introduces simulated errors in the corrected data and then performs correction again. Using statistics such as the number of corrections performed on the raw data, and the number of missed simulated errors, the quality of the error detection is estimated.

First, using a set of sequenced known data we estimate the pattern in which errors occur in the data, and store it as what we will call an error profile. Then, the following procedure is applied: The raw data (0*e*) is corrected with both approaches producing two corrected datasets (1). Artificial errors are introduced in both corrected sets, producing simulated error sets (1*e*). Each of them is corrected with both error correction approaches, producing doubly-corrected sets (2). Statistics are collected to evaluate the two approaches against each other as it is seen in Fig. 3.

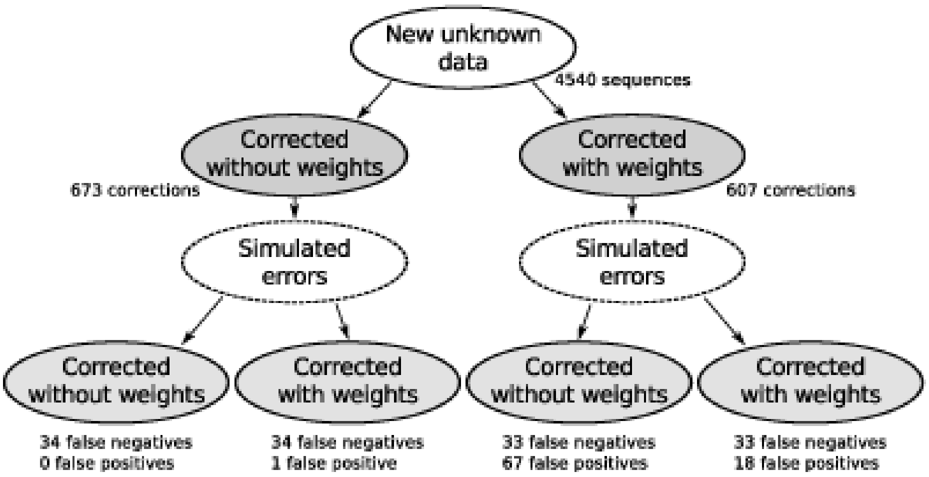


Fig. 3. Repeated application validation approach

The amount of missed simulated errors when (1*e*) was corrected to (2) should be roughly the same as the amount of missed real errors (false negatives), because of our error profile. If the "improved" correction approach leads to the same number of false negatives yet makes less corrections overall, this is a clear sign that the number of false positives has also decreased.

2.6 Metagenomic Workflow Software Library

To data processing pipeline used to perform the experiments with the analytical approach is executed using a simple library that interfaces with external tools and incorporates the error detection and validation.

The software is presently being reworked into a more flexible workflow system that can be extended to execute arbitrary programmable parametric workflows and can solve a wider variaty of tasks, as such would be very significant contribution to the Bioinformatics community. A future mature version of that workflow library should be able to execute workflows for a wide variety of tasks. The software is described in more detail in [4].

3 Results and Discussion

3.1 Experimental Results

On a test run of 4540 sequences, we performed corrections with both the similarity-weighted ("improved") approach and the naïve ("base") approach. The naïve approach produced 673 corrections, while the similarity-based produced only 607, or 66 less, which is a 10% decrease in the number of corrections.

After simulated errors were introduced using our estimated error profile, the error correction approaches again. As illustrated in Fig. 3 in the set initially corrected by the naïve approach, the false negatives were 34 for both, while in the set initially corrected by the similarity-based approach, the false negatives were 33 for both. In other words, the number of false negatives has remained almost the same, and has even seen a statistically insignificant decrease.

The result is consistent with the expectation for a decrease in the number of spurious corrections. Using the similarity-weighted approach, the number of corrections has decreased with 10%, yet the number of correctly identified simulated errors has not decreased, which suggests that the decrease in corrections is at the expense of false positives. Experiments with the remaining clusters extracted from the input data yielded similar but, due to their smaller size, less significant results.

3.2 Software

The computational experiments are executed with a software module written in Python and Cython. The module interfaces with arbitrary external aligning and clustering software using the FASTA and ACE formats, including MUSCLE [9], MAFFT [8] and CD-HIT-454 [7]. A custom aligner combining CD-HIT with MAFFT or MUSCLE for a better quality to CPU time ratio is also included. For neural networks, the Python FANN library is being utilised [12].

3.3 Further Work

The software is presently being rewritten using the Twisted networking framework [13] which will replace any ad-hoc asynchronous code and will allow distributing computing tasks and interfacing with network-aware tools. The new version utilises programmable workflow descriptions written in the YAML language that provide flexibility in constructing pipelines, workflows and various experiments.

The neural network training and streamlined validation procedures are being incorporated within this new mini-framework for workflows. This will allow for further extensive testing of the presented analytical approach, along with the training of the neural network, and will also allow for experimentation with modified network designs and inputs.

4 Conclusions

The suggested similarity weight used in the proposed analytical error correction approach shows results consistent with an improvement in the quality of error correction during the presented experimental validation, and can be predicted to flag 10% less errors without missing actual errors in the process.

The workflow library in development would not only allow the execution of more validation experiments and training of the neural network, but will also be a useful contribution to the Bioinformatics community where custom workflows are often necessary for data analysis once extended to cover a wider variety of tasks.

The nature of the data – largely inter-related yet not well explored and chaotic – makes it suitable to apply the inference capability of artificial neural networks in searching for a complex formulaic between the input and the occurrence of errors.

Acknowledgements

This work was supported by the European Social Fund through the Human Resource Development Operational Programme under contract BG051PO001-3.3.06-0052 (2012/2014).

References

1. Nelson, K., White, B.: Metagenomics and its applications to the study of the human microbiome. Metagenomics: Theory, Methods and Applications (2010) 171–182
2. Kristensen, D., Mushegian, A., Dolja, V., Koonin, E.: New dimensions of the virus world discovered through metagenomics. Trends in Microbiology **18**(1) (2010) 11–19 (doi:10.1016/j.tim.2009.11.003, URL `http://www.biomedsearch.com/nih/New-dimensions-virus-worlddiscovered/19942437.html`

3. Valverde, J.R., Mellado, R.P.: Analysis of metagenomic data containing high biodiversity levels. PLoS ONE **8**(3) (2013) e58118 (doi:10.1371/journal.pone.0058118 URL `http://www.plosone.org/article/info:doi/10.1371/journal.pone.0058118`)
4. Krachunov, K., Vassilev, D.: An approach to a metagenomic data processing workflow. Journal of Computational Science **4** (2013) (in press)
5. Weisburg, W.G., Barns, S.M., Pelletier, D.A., Lane D.J.: 16S ribosomal DNA amplification for phylogenetic study. J. Bacteriol. **173**(2) (1991) 697–703
6. Mande, S., Mohammed, M., Ghosh, T.: Classification of metagenomic sequences: methods and challenges. Briefings in Bioinformatics **13**(6) (2012) 669–681 (doi:10.1093/bib/bbs054 URL `http://bib.oxfordjournals.org/content/early/2012/09/07/bib.bbs054.full.pdf`)
7. Li, W., Godzik, A.: Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics **22**(13) (2006) 1658–1659 (doi:10.1093/bioinformatics/btl158 URL `http://bioinformatics.oxfordjournals.org/content/22/13/1658.full.pdf`)
8. Katoh, K., Kuma, K., Toh, H., Miyata, T.: MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleid Acid Research **33**(2) (2005) 511–518
9. Edgar, R.: MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics **5**(1) (2004) 113 (doi:10.1186/1471-2105-5-113)
10. Krachunov, M., Petrov, P., Popov, I., Vassilev, D.: Computational challenges in a metagenomics processing pipeline. In: Proceedings of the 6th International Conference on Information Systems & Grid Technologies (ISGT), Sofia (2012) 302–311
11. Zerbino, D., Birney, E.: Velvet: algorithms for de novo short read assembly using de bruijn graphs. Genome Res. **18**(5) (2008) 821–829 (doi:10.1101/gr.074492.107)
12. Nissen, S., Spilca, A., Zabot, A., Morelli, D., Nemerson, E., Freegoldbar, Megidish, G., Joshwah, M., Pereira, M., Vogt, S., Hauberg, S., Leibovici, T., Massa, V.: Fast artificial neural network library (2006) (URL `http://leenissen.dk/fann/`, accessed 4 May 2013)
13. Zadka, M., Lefkowitz, G.: The twisted network framework. 10th International Python Conference (2002) (URL `http://www.python.org/workshops/2002-02/papers/09/`, accessed 12 August 2013)

URL https://twistedmatrix.com/users/glyph/ipc10/paper.html

NoSQL Solutions to Handle Big Data

Emanuela Mitreva[1] and Kalinka Kaloyanova[2]

Faculty of Mathematics and Informatics, Sofia University "St. Kliment Ohridski",
5, James Bourchier Blvd., 1164 Sofia, Bulgaria
[1]emitreva@gmail.com, [2]kkaloyanova@fmi.uni-sofia.bg

Abstract. The growth of data nowadays raises a question how it can be processed effectively, based on new trends in IT area. It placed a need for concepts, methods, technologies and tools, with which the large amount of generated data will be handled and also transformed into knowledge and value for the business. This paper presents NoSQL solutions, describes their main characteristics and discusses how they can be used as tools for handling big data.

Keywords: NoSQL, big data, key-value stores, column-based databases, document databases, graph databases

1 Introduction

Big data refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze. This definition is intentionally subjective as it is assumed that, as technology advances over time, the size of datasets that qualify as big data will also increase [1].

The growth of the data is just one side of the problem with big data; the other is the necessity to store not only structured data, but pictures, video, files and unstructured data. A very indicative example is that the relational model can neither handle the traffic social sites like Facebook and Twitter generate, nor the type of data they want to store. RDBMS cannot handle properly an amount of data larger than the amount, stored on a server. Some RDBMS offer limited possibility to scale vertically or horizontally, but it is difficult to make the data consistent, while changing it on several servers, thus making the queries much heavier than they can be if the data is just on one server. The difficulties, described above, show a need for alternative solutions, which will handle big data more effectively than the RDBMS solutions and in the next part of this article we will present one approach to handle large amounts of data.

2 NoSQL Basic Concepts

There is neither strict definition nor any standard for what the term *NoSQL* means. However the experts agree upon one thing – NoSQL means Not Only SQL.

Generally, any database that is not RDBMS, upholds schema-less structures, does not necessary comply with the ACID properties (atomicity, consistency, isolation, and durability) and promises high availability and support for large data sets in horizontally scaled environments can be categorized as a *NoSQL data store* [2]. Some argue that NonRel is a more appropriate term than NoSQL.

Having a schema-less structure can be quite beneficial, especially in case of unstructured data, or when the records are not with fix content. What RDBMS is offering is fixed structure of the data – in columns and rows – however not all the data, generated nowadays, can fit into this model. By using the relational model it is possible to change the structure of a table – to add or delete columns, however in very big tables, this can be really heavy operation, with significant down-time and down-time is something customers want to minimized.

2.1 Base vs. ACID

The relational model complies with the ACID properties. Databases with ACID properties have more strict rules for the data and the data consistency. These properties require data to be in consistent state at any time, which will take a lot of time for heavy operations, even reading only. Therefore the relational model is not suitable for applications with huge traffic.

Most NoSQL databases do not support the ACID (Atomicity, Consistency, Isolation, Durability) properties, but properties less strict than the ACID ones, which are called BASE. They consist of three principles – Basic Availability, Soft state and Eventual consistency. Basic availability means that the data is available at any time, what is not ensured, is that the data is consistent at every node (usually the availability of the data in NoSQL solutions is made possible by having the same data on more than one node). Moreover some of the logic for keeping the data in consistent state is moved to the application and the developer is the one that should ensure the consistency of the data. The other principal of BASE is eventual consistency, which states that the data is not consistent at all time, but will be at some time. Any changes are populated sooner or later on all the copies. For most applications this is enough and there is no need of the limitation of the data to be consistent at any time.

2.2 CAP Theorem

All systems follow the CAP theorem, which stands for Consistency, Availability, and Partition Tolerance. In [3] these properties are described as follows: Consistency – *all clients always have the same view of the data*, Availability - *each client can always read and write*, and Partition Tolerance – *the system works well despite physical network partitions*. For all of the database storage, they could take only two of those characteristics. Existing RDBMS takes consistency and availability but cannot offer partition tolerance. On the other hand the NoSQL takes partition tolerance with giving up either consistency or availability.

Achieving durability has long been the bottleneck for database systems. It is easy to understand why writing to disk slows down the database. That is the reason NoSQL chooses availability over consistency – instead NoSQL support eventual consistency. But the CAP theorem offers choices that lead to trade-offs – speed over durability [4].

2.3 Map-Reduce

The fundamental concept of NoSQL databases is the underlying Map/Reduce approach, which is a framework that supports the handling of large data volumes over distributed network nodes. To handle very large data volumes and operations on them, any problem is divided into sub-problems that are distributed to other network nodes – the map step. Subsequently, these network nodes are allowed to do the same distribution process. In the end the sub-problems are solved by certain network nodes and the results are passed on to the master node that interprets the results – reduce step [5].

2.4 Scalability

According to Cattel [6] the NoSQL solutions offers:

- the ability to horizontally scale “simple operation” throughput over many servers;
- the ability to replicate and to distribute (partition) data over many servers.

Steps to make a relation database more scalable are taken, however there are some limitations to it. For a query to benefit from a scalability of the relation database, its resulting subset should consist of only records on the same server. Having results from multiple servers leads to more time to generate the result, therefore losing the efficiency of a truly scalable database as the NoSQL solutions can be.

3 NoSQL Solutions

NoSQL growing popularity leads to an impressible number of NoSQL databases. Based on the data and access model, Cattel [6] differentiate several types:

- Key-Value Stores;
- Document Stores;
- Column Stores/Extensible Record Stores;
- Graph Stores.

3.1 Key-Value Stores

Data and Access Model. Data is usually consisting of a string, which represents the key in the *key-value* relationship. The data itself can be either some kind of primitive of the programming language (a string, an integer, an array) or some unstructured data. This replaces the need for fixed data model and makes the requirement for properly formatted data less strict [7].

Indexing and Searching. The big advantage of key-value stores is in their simplicity – all records are identified only by a key and the value can be in any format – either structured or unstructured data. Every record has unique key so the data store provides fast, indexes lookups. They allow simple read and write operations and the objects are stored as binary ones, identified by the key. The fundamental weaknesses for those kinds of databases is that they offer only a search by the key and do no support indexes on anything other then the key, i.e. key-value stores do not support secondary indexes.

Joins. Join operation is not supported by key-value stores, but this can be done by the application. Another function of the application is to manage the serialization and deserialization of the data, meaning the value can consist of any structured or unstructured data, e.g. fixed records or complex structures, but it should be parsed by the application itself in order to be used (if any parsing is needed).

Usage. Key-value stores are one of the simplest NoSQL solutions. It does not offer much flexibility of the data or options for searching, but if the data, that should be stored, can be presented in key-value pairs and there is a need to search only by a key, key-value store is the best option for it.

Examples. Prominent examples of key-value stores are Amazon's Dynamo, Oracle NoSQL and CouchDB.

3.2 Document Stores

Data and Access Model. Document oriented databases are used to store, manage and retrieve the structured or semi-structured data in the form of a document. The main concept in these types of databases is *document* – similar to a record in relational databases, but distinguished in many aspects such as it is less rigid and use different format to store data. The document oriented databases store data in JSON, BSON or XML format and many others [3]. The main difference from the key-value store models is that document stores offers practically limitless ability to nest elements.

Indexing and Searching. Unlike the key-value stores, these systems generally support secondary indexes and multiple types of documents (objects) per database, and nested documents or lists [6].

Examples. The most notable examples of a document NoSQL database is MongoDB.

3.3 Extensible Records Stores (Column-Based)

Data and Access Model. Extensible records can be row-based or column-based. The row-based stores rows are split across nodes through sharding on the primary key. They typically split by a range so the result is that queries on ranges of values do not have to go to every node [6].

In Column-Stores each database table stores each column separately, with attribute values belonging to the same column [8].

The column groups must be pre-defined with the extensible record stores. However, that is not a big constraint, as new attributes can be defined at any time. Rows are analogous to documents: it is possible to have a variable number of attributes, the attribute names must be unique, rows are grouped into collections (tables), and an individual row's attributes can be of any type [6].

Indexing and Searching. The indexing in a row-based store is by the rowid, and the indexing in the column-based store is done on the data and it is pointing to the rowid. So the searching in column-based stores for records matching a single column is quite fast and can be done by a single operation. However queries, when the record should consist of a several columns, are much slower and heavier.

Usage. Column-based stores are more effective, when most of the used queries extract only part of the columns. They are mainly used in OLAP (OnLine Analytical Processing) and Data Mining operations [8].

Examples. Vertica and C-Store are examples of column-based databases.

3.4 Graph Stores

The graph is one of the fundamental mathematical abstractions in computer science and it is considerably used - mostly while representing hierarchy and relations, especially when the depth of the hierarchy is more than two. RDBMS offers limited or no ability to use hierarchical queries and to represent hierarchical data, e.g. Oracle supports hierarchical queries, but in any other RDBMS there is short to none support for such.

Data and Access Model. A graph database is a storage engine that is specialized in storing and retrieving vast networks of data. It efficiently stores nodes and relationships and allows a high performance traversal of those structures. Properties can be added to nodes and relationships [9].

Graph databases are well suitable for storing numerous kinds of domain models. In most other modeling approaches, the relationships are reduced to a single link without any identifier and attributes. Graph databases allow one to keep the rich relationships that originate from the domain, equally well represented in the database without resorting to also modeling the relationships as *things*. There is very little *impedance mismatch* when putting real-life domains into a graph database [9].

Usage. The graph representation of data can be quite useful for the social sites – Facebook, Twitter, LinkedIn, etc., which are generating big part of the traffic on the internet.

Examples. Neo4j is one of the most used graph databases. It supports the ACID properties. Neo4j stores data structured as graphs. Each graph consists of nodes, connected by relationships. It allows high query performance on complex data, while remaining intuitive and simple for the developer [9].

4 Comparison between the NoSQL Types of Databases

The NoSQL solutions are quite diverse in the functionalities and basic properties, which they provide and this makes them suitable in different situations. The relational databases are optimized and are best for handling transactions. The NoSQL solutions on the other hand provide good properties to handle unstructured data (video, pictures, rows with differing number of properties).

Some databases are optimized for read operations (Neo4j), others like Cassandra and BigTable, which offers consistency, high availability, partition tolerance, and persistence, are more suitable for frequent write operations (i.e. making them suitable for logging). In almost every area of business, e.g. telecommunication, there is a lot of information, which is good to be logged in case of failure or requests for some details for specific operation from the user, making the logging necessary, but not frequently used. The large amounts of logging information cannot be handled so well with the relational model, especially if the structure of the table needs to be changed over time.

The NoSQL databases do not usually conform to the ACID properties, but to the BASE ones, thus making them unsuitable for transactions, e.g. bank transactions. Still there are some of them (DynamoDB, CouchBase, Neo4j), which do comply with ACID and as a result these DBs offer better consistency in that way they possess the advantages of the relational databases, i.e. consistency and still offer better scalability.

Also most NoSQL solutions allow indexing – either only primary key (most key-value stores) or primary and some secondary indexing and the use of Map/ Reduce. The indexing provides the user with better searching on the fields on which the index is created. If only primary indexing is required, then key-value store is the best solution for these needs.

Unlike RDBMS, the NoSQL solutions do not have a common query language – SQL. They rather have APIs or even SQL-like languages (GQL, CQL) to manage the database and query its data. By providing access to the data via APIs, they are handling better the operations and changes on the data. Moreover nowadays most companies do not provide direct access to their data, but via

Table 1. Comparison of several NoSQL databases

	DynamoDB	CouchBase	MongoDB	Big Table	Cassandra	Neo4j
Type	Key-value	Document	Document	Column	Column	Graph
Developer	Amazon	Couchbase	10gen	Google	Apache	Neo Technology
Language	Java	C/C++	C++	C/C++	Java	Java
Indexing	Single or composite hash-range key	Two-dimensional	Single or multiple fields	Indexes on ranges, not values	On data, not on attribute	Nodes & relationships keys and relationships
Secondary index	No	Yes	Yes	Yes	Yes	Yes
Map/Reduce	Yes	Yes	Yes	Yes	Yes	Yes
BASE or ACID	ACID	ACID	BASE	BASE	BASE	ACID
Concurrency Control	ACID	ACID, no transactions	Locks	Locks	MVCC	Transactions in the Java API
Protocol	Put and get APIs	JSON	BSON	APIs	Custom, binary (Thrift)	HTTP/REST (or embedding in Java)
Storage	SSD	Disk	Memory mapped b-trees	GFS	Memtable	Disk
Queries	Key-value GET/PUT operations using a user-defined PK	Spatial queries	JavaScript expressions	GQL	CQL	Pattern-matching-based query language (*Cypher*)

Characteristics	Consistency;
	High Availability;
	Eventually Consistent Reads
		Consistency;
		High Availability;
		Persistence
			Consistency;
			Partition;
			Tolerance;
			Persistence
				Consistency;
				High Availability;
				Partition Tolerance;
				Persistence
					High Availability;
					Partition Tolerance;
					Persistence
						Optimized for reads

Best of use	Large to big db solution	Session, user profile, content stores	Dynamic queries, frequently written, rarely read statistical data	Scale (hundreds to thousands machines)	Write often, read less	Complex data relationships and queries

some APIs, which they have to develop, in the case with NoSQL, they directly provide APIs.

Table 1 presents a summary of the most popular NoSQL databases [10–14].

5 Conclusion

In this article we have presented NoSQL databases as a possible solution to handle big data. We have described and compare the different types of NoSQL – key-value, document, column-based and graph stores. The comparison shows that the different NoSQL stores can be used in different cases, based on the need of the customer – whether he is using it for mostly reads and less writes or the opposite. What most NoSQL databases offer is better scalability, optimizations for either reads or writes, even some comply with the ACID properties. So based on the need to be met and the characteristics that are expected to have, the diversity of NoSQL databases can offer effective solutions to manage big data.

Acknowledgments

This paper is supported by the Scientific Research Fund of the Sofia University "St. Kliment Ohridski" under contract No. 43/2013 and the European Social Fund through the Human Resource Development Operational Programme under contract BG051PO001-3.3.06-0052 (2012/2014).

References

1. McKinsey Global Institute: Big data?: The next frontier for innovation, competition, and productivity (2011) `http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation`
2. Tiwari, S.: Using Oracle Berkeley DB as a NoSQL Data Store (2011) `http://www.oracle.com/technetwork/articles/cloudcomp/berkeleydb-nosql-323570.html`
3. Padhy, R.P., Patra, M.R., Satapathy, S.C.: RDBMS to NoSQL: Reviewing Some Next-Generation Non-Relational Database's. International Journal of Advanced Engineering Sciences and Technologies **11**(1) (2011) 15–30
4. Burd, G.: NoSQL sysadmin. Login **36**(5) (2011) 5–12
5. Scholz, J.: Coping with Dynamic, Unstructured Data Sets – NoSQL a Buzzword or a Savior? In: Schrenk, M., Popovich, V.V., Zeile P. (eds.). Proceedings of 16th International Conference on Urban Planning, Regional Development and Information Society REAL CORP (2011) 121–129 ISBN: 978-3-9503110-1-3
6. Cattell, R.: Scalable SQL and NoSQL Data Stores. ACM SIGMOD Record **39** (2010) 12–27
7. Seeger, M.: Key-Value stores: a practical overview. Media (2009) 1–21
8. Kaur, J., Kaur H., Kaur, K.: A Review on Document Oriented and Column Oriented Databases. International Journal of Computer Trends and Technology **4**(3) (2013) 338–344

9. Spring Framework: Chapter 17. Introduction to Neo4j. `http://static.springsource.org/spring-data/data-graph/docs/current/reference/multi/neo4j.html`
10. Best NoSQL Databases. Unbiased, Data-Driven Comparisons. `http://nosql.findthebest.com/`
11. Comparison of Popular NoSql databases (MongoDb, CouchDb, Hbase, Neo4j, Cassandra). `http://mad4nosql.blogspot.de/2012/12/comparison-of-popular-nosql-databases.html`
12. NOSQL Databases `http://nosql-database.org/`
13. Amazon DynamoDB `http://aws.amazon.com/dynamodb/faqs`
14. MongoDB Manual `http://docs.mongodb.org/manual/`

Cloud Computing – Security Concerns and Countermeasures

Radoslav Ivanov

Faculty of Mathematics and Informatics, Sofia University “St. Kliment Ohridski”,
5, James Bourchier Blvd., 1164 Sofia, Bulgaria
radoslav.h.ivanov@gmail.com

Abstract. Cloud based services are subject to increased popularity in recent years due to number of benefits provided by cloud environments including scalability, on demand access to resources, flexibility, efficient utilization of resources and decreased costs. Cloud computing brings new opportunities but also introduces new challenges. Securing data and services in the cloud is one of them and is responsibility of both cloud providers and consumers. They should understand the potential risks and countermeasures and apply proper mitigation techniques. This paper will explore the various security concerns related to cloud computing and techniques for addressing them.

Keywords: cloud computing, security, cloud security

1 Introduction

Cloud computing is growing domain that draws attention of both industry and academy in recent years. Thanks to utilization of virtualization and automation technologies, cloud computing is providing scalable and on demand access to resources. It increases efficiency of data centers and decreases costs due to ability of flexible resource allocation and sharing of physical resources between users. Infrastructure and services are hosted in data centers and available to users via network or internet.

While cloud computing is attractive due to the number of benefits it provides, it also raises some privacy, security and legal concerns regarding data and services hosted in the cloud. These should be properly addressed to decrease risks and ensure smooth operations.

This paper will explore the security concerns related to cloud computing and the key challenges associated with different cloud models. Furthermore it will discuss various options for securing cloud services and infrastructure and the shared security responsibilities between provider and consumer.

2 Cloud Computing Models

Each cloud computing model faces different levels of risk and challenges vary depending on which model is used. Following is brief overview of different cloud models.

2.1 Cloud Deployment Models

There are several approaches in adopting cloud services depending on the needs and abilities of the organizations.

Public Cloud. Public clouds are available to general public and various vendors provide public cloud environments where services are offered for free, on pay-per-use or subscription basis. Cloud service providers are hosting the infrastructure and services in their own data centers and make them accessible via internet. Sharing the resources in such multi-tenant environment improves the utilization rates and can reduce costs significantly. Consumers rent the resources instead of investing in their own infrastructure and they can request and release resources according to their needs.

Private Cloud. Private cloud is cloud environment that is used by single organization. It uses the same principles that are behind public clouds and offers many of the benefits of cloud computing like scalability, efficiency, dynamic allocation of resources and so on, but resources are not shared with external entities. It provides maximum control over resources, but requires larger investments and may lack agility when additional capacity is required.

Community Cloud. Community cloud is cloud environment that is shared between several organizations and is used solely by them. It may be owned and managed by one or more of the participating organizations or may be outsourced to third parties.

Hybrid Cloud. Cloud infrastructure is a composition of two or more distinct cloud infrastructures (private, community or public) that remain unique entities but are bound together [1]. Hybrid clouds allow combined usage of private and public clouds depending on nature of the tasks, which is helping to optimize security and privacy while still maintaining balance between costs and scalability.

2.2 Cloud Service Models

In addition to the deployment models there are also different service models. There are three main service models in cloud computing – Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS) (see Fig. 1).

Infrastructure as a Service (IaaS). Infrastructure like storage, processing, network and other computing resources is delivered as a service. Consumers can deploy and run their own software and have full control on it, but they do not have control over underlying physical infrastructure.

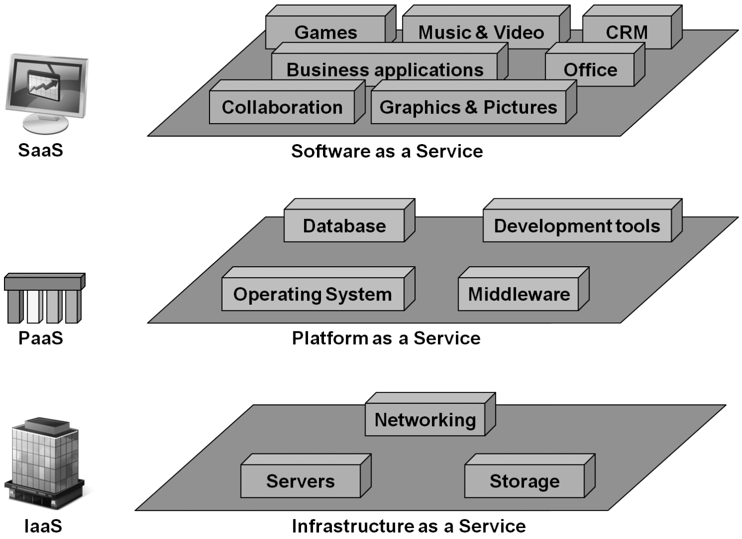


Fig. 1. Cloud service models

Platform as a Service (PaaS). Operating systems, middleware services, back-end technologies or combination of them are delivered as a service and consumers can create and deploy applications using these services. Consumers have control over their applications but have no control over underlying infrastructure, operating systems and middleware.

Software as a Service (SaaS). Enables software applications hosted in provider's cloud environment to be delivered as a service. Consumers can use provided applications, but do not have control over them and underlying services and infrastructure. They might be granted access to limited configuration options.

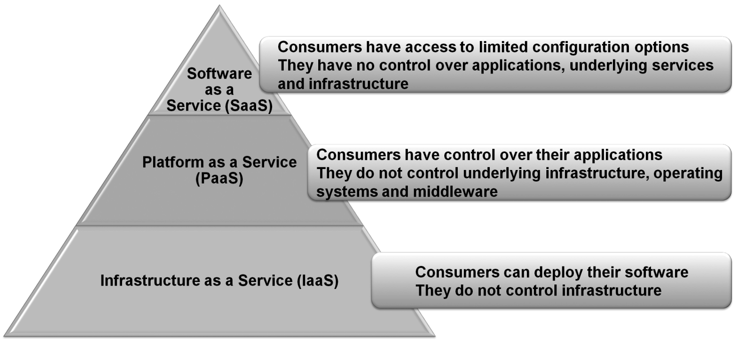

Software as a Service (SaaS)
Consumers have access to limited configuration options
They have no control over applications, underlying services and infrastructure
Platform as a Service (PaaS)
Consumers have control over their applications
They do not control underlying infrastructure, operating systems and middleware
Infrastructure as a Service (IaaS)
Consumers can deploy their software
They do not control infrastructure

Fig. 2. Consumer control

3 Problem Description

Cloud computing provides the opportunity of outsourcing operations to cloud providers and thus shifting the responsibility of maintaining infrastructure or services to the provider. Although control over data and operations is shifted to the provider, it does not necessarily mean that all responsibility about security is shifted to the provider as well.

For example in the IaaS model, it is cloud provider's responsibility to provide the client pre-configured virtual machines that are up to date with applied latest security patches. It is also provider responsibility to secure data center facilities and infrastructure. Once the client takes control over provisioned virtual machines, it becomes client responsibility to maintain their patch level after the initial deployment. If the client does not apply regularly security patches, then security may be compromised.

This is why cloud security is joined responsibility of providers and consumers of cloud services. To ensure high security levels, organizations should understand the risks and examine how providers implement and manage security on their behalf. This is critical since although part of the security duties are shifted to the provider, the final responsibility for compliance and protection of critical assets is still in customer's hands.

4 Security Concerns

Following sections will review some of the main security concerns related to cloud computing and mechanisms that can be applied to mitigate the risks.

4.1 Data Access and Control

Moving data and applications to the cloud brings concerns regarding privacy and security of data. Loss of data in case of service outage or malicious deletion could also have serious consequences. While cloud providers usually have security measures and recovery mechanisms deployed in their facilities, it is important to evaluate their security practices before moving your data to particular provider.

4.2 Multi-tenancy Issues

In cloud computing resources and services are provided to multiple tenants that share services or infrastructure like memory, disk partitions, CPU caches and other components. Even though there is a virtualization layer separating physical resources from guest operating systems, there is a concern that attackers could eventually gain access to underlying platform. Potential compromise of the hypervisor can lead to compromise of all shared resources under its control.

4.3 Account or Service Hijacking

Stolen credentials could provide attackers with access to sensitive data and critical resources. This could lead to data loss or manipulation and usage of resources for criminal activities.

4.4 Vendor Lock-in

Once users move data and applications in the cloud it may become difficult to move away to another provider. To reduce this risk, possibilities for extracting data from the cloud and migration to other providers should be evaluated in advance before selecting a provider.

4.5 Regulatory Compliance

Regulatory authorities may have certain requirements for organizations dealing with sensitive data like health records, financial or personal data. In cloud environments, it is often difficult to locate where data is stored and this may result in security and compliance issues that prevent the usage of public clouds in certain cases.

This could be resolved by using private cloud or using provider that allows selection of physical location where data is stored. When moving to the cloud, organizations should investigate whether the cloud provider meets their requirements for compliance with the regulations and ensure that formal contracts are in place to guarantee that these requirements are met.

4.6 Reliability

Cloud providers usually provide certain availability guarantees and recovery options backed up by service level agreements. When moving mission critical services to the cloud, customers should carefully review service level agreements of the providers and plan accordingly their mitigation strategy to address outages.

5 Security Countermeasures

Following section will review some of the security countermeasures that could be applied to mitigate the risks and increase security in cloud environments. These measures cover multiple layers and levels of defense starting from physical facilities and going through hardware, software, application and data security, that should be considered during security planning and implementation process.

The section will provide brief guidelines, while more details and deeper insights may be found when following the references and the resources mentioned in them.

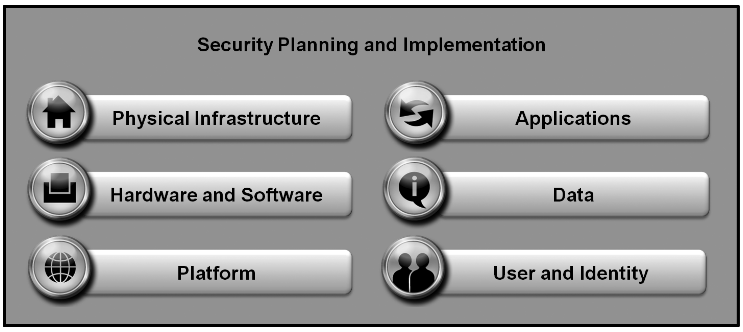


Fig. 3. Security considerations

5.1 Physical Infrastructure

Physical cloud infrastructure like servers, storage, network equipment and other components must be secured to prevent physical unauthorized access. Cloud facilities should have proper security control and monitoring of physical access. This may include video surveillance, security guards, biometric and card security, alarm systems, fire protection and other safety and access control mechanisms.

5.2 Hardware and Software

Resource provisioning environment should be properly secured as well. This includes hypervisor, underlying hardware infrastructure and physical network.

When image catalogs are provided, images must be secured and properly protected from corruption and abuse. Many clients expect these images to be cryptographically certified and protected [2]. Consumers should regularly apply security patches to the provisioned images, maintain security of applications they deploy and implement security policies that cover their needs.

Cloud environment provides to consumers virtualized resources, while the actual physical resources are shared among tenants. It is necessary to ensure that tenant domains are properly isolated.

Intrusion detection and prevention mechanisms should be in place to ensure early detection of potential threats and proper reaction for minimizing impact of malicious attacks.

5.3 Platform

Securing the hardware and software is a prerequisite for cloud security, but the overall platform should be secured as well. This includes securing servers, client devices, communication channels and application programming interfaces (APIs). It requires proper authentication, authorization and identity management mechanisms.

5.4 Applications

All security requirements of traditional applications should be applied also to cloud applications. Cloud application development should follow secure development process and respect security guidelines and practices.

5.5 Data Protection

All sensitive data stored in the cloud should be properly protected, including archived data. This may require encryption of data and management of encryption keys. Encryption should be applied not only to sensitive data stored in the cloud but also when data is transferred from and to the cloud environment, especially data stored or moving in and out of public clouds. Depending on data sensitivity it may be necessary to explicitly define how data will be encrypted and archived, what mechanisms will be used to prevent data loss and who has access to it.

When data is subject to regulatory or legal restrictions, organization must perform deep analysis of all requirements prior to cloud deployment to make sure it can maintain the necessary level of control.

5.6 User and Identity Management

Organizations can adopt strong identity management mechanisms like strong password requirements, two factor authentication and regular monitoring of user activities.

5.7 Security Planning and Implementation

Security should be integrated in the whole process of building, delivery and usage of cloud services. Adding security on top of already existing systems, built without security in mind, is hard and expensive initiative with questionable results. Organizations should integrate security early in the overall cloud planning process. This requires identification of security objective, potential threats and countermeasures.

Various vendors and organizations are providing guidelines and best practices for enabling security in cloud environments. Papers like IBM's Security and high availability in cloud computing environments [2] and Intel's Cloud Computing Security Planning Guide [3] review some of the security challenges of cloud computing, provide guidelines and vendor specific solutions for them. Other organizations like Cloud security alliance (CSA) are providing more general guidelines. CSA report about top cloud computing threats in 2013 [4] provides information helping to understand some of the most common security threats and can assist organizations in taking informed decisions about cloud adoption and related security threats.

6 Conclusion

Cloud computing offers various advantages like scalability, efficiency and cost-effectiveness. Different cloud computing models are providing flexibility for delivering services but shifting control over data and services to third parties raises a number of security concerns.

When using public cloud services to outsource operations, organizations should carefully review written business agreements and investigate security measures and responsibilities covered by the cloud provider.

Private clouds provide many of the benefits of cloud computing while eliminating some of the security concerns related to public cloud environments. Since data and services stay inside the organization problems like vendor lock-in and regulatory compliance are minimized. On other hand, private clouds require significant investments and efforts due to need of purchasing and maintaining own infrastructure and cloud environment.

Security is shared responsibility between cloud providers and consumers. Each of them is responsible for securing the assets under their control.

This paper reviews the main security concerns related to cloud computing, the challenges for cloud providers and consumers and various options to deal with them.

Acknowledgments

This work was supported by the Project DDVU 02/22/2011 funded by the Bulgarian Science Fund.

References

1. Mell, P., Grance, T.: The NIST Definition of Cloud Computing. National Institute of Standards and Technology, U.S. Department of Commerce. Special Publication 800-145 (2011). http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf
2. IBM Global Technology Services: Security and high availability in cloud computing environments. Technical White Paper (2011) http://www-935.ibm.com/services/za/gts/cloud/Security_and_high_availability_in_cloud_computing_environments.pdf
3. Intel IT Center: Cloud Computing Security Planning Guide (2012). http://www.intel.com/content/dam/www/public/us/en/documents/guides/cloud-computing-security-planning-guide2.pdf
4. Cloud Security Alliance: The Notorious Nine: Cloud Computing Top Threats in 2013 (2013). https://cloudsecurityalliance.org/media/news/ca-warns-providers-of-the-notorious-nine-cloud-computing-top-threats-in-2013/

Comparative Analysis of Brain Data Clustering

Sergey Milanov[1], Olga Georgieva[1], and Petia Georgieva[2]

[1] Faculty of Mathematics and Informatics, Sofia University "St. Kliment Ohridski",
5, James Bourchier Blvd., 1164 Sofia, Bulgaria
s_milanov@yahoo.com, o.georgieva@fmi.uni-sofia.bg
[2] Signal Processing Lab, IEETA, University of Aveiro, Portugal
petia@ua.pt

Abstract. The goal of this study is to implement different clustering data mining methods as a tool to analyze brain data (Electroencephalogram signals) in order to discriminate positive and negative human emotions while subjects were exposed to external stimulus (images). The results of the clustering are compared with respect to methods' reliability to correctly detect the two EEG clusters associated with positive and negative emotions respectively. The clustering procedure is preceded by a preprocessing step of feature selection that improved the clustering results.

Keywords: data mining, cluster analysis, biosignal retrieval, EEG signals

1 Introduction

The currently investigated and established methods of reliable diagnostic and cure decisions are widely supported by the new generation of information and communication devices. This new technological base has to be suitably supplied by appropriate algorithm for data processing and analysis. The information retrieved from biosignal data needs correct explanations to assess the states and to propose timely decisions [10].

Being some of the most popular data mining algorithms [1, 2], clustering techniques have been extensively used in bioinformatics to analyze biomedical data. The objective of the present paper is to find a reliable clustering method(s) able to discriminate human emotions based on Electroencephalogram (EEG) signals. The EEG data was collected while subjects were exposed to images which typically provoke positive and negative emotions. Different clustering methods are compared with respect to their reliability to correctly detect the two EEG clusters associated with positive and negative emotions respectively. The clustering procedure is preceded by a preprocessing step of feature selection that improved the clustering results.

The paper is organized as follows: Section 2 describes the acquisition process of the EEG signals and the features that constitute the dataset. The overview of feature selection methods and theoretical background of cluster analyses is presented in Sect. 3. Section 4 describes the experimental framework and analyses results.

2 Data Set Description

The objective of the tests performed was to distinguish emotional bio-signals evoked by viewing selected affective pictures. During the experiments, the EEG signals were recorded. A total of 26 female volunteers participated in the study, 21 channels of EEG positioned according to the 10–20 system and 2 EOG channels (vertical and horizontal) were sampled at 1000 Hz and stored. The recorded channels were Frontal and Parietal lobe (FP channels), Frontal lobe (F channels); Temporal lobe (T channels); Central lobe (C channels), Parietal lobe (P channels) and Occipital lobe (O channels). The signals were recorded while the volunteers were viewing pictures selected from the International Affective Picture System (IAPS). A total of 24 of high arousal (> 6) images with positive valence (7.29 ± 0.65) and negative valence (1.47 ± 0.24) were selected. Each image was presented three times in a pseudo-random order and each trial lasted 3500 ms: during the first 750 ms, a fixation crosses was presented, then one of the images during 500 ms and at last a black screen during the 2250 ms.

First, the row EEG signals were preprocessed. They are filtered, eye-movement corrected, baseline compensation and epoched using NeuroScan equipment. The ensemble average for each condition was also computed and filtered using a Butterworth filter of fourth order with passband (0.5–15 Hz). Twelve features are stored corresponding to the latency (time of occurrence) and the amplitude of the first three maximums and three minimums observed in the ensemble averaged EEG signals. The last column of the data set is the class that corresponds to the positive (Class_1) or negative (Class_2) image.

3 Theoretical Background

In order to discover specific knowledge in the Visually Evoked Potentials (VEP), these are EEG signals collected during visual stimuli, several clustering algorithms were applied to the same EEG dataset. The predictive accuracy of the clustering algorithm is improved by a preliminary step of feature selection.

3.1 Feature Selection

In applying the clustering algorithms the following criteria for feature dimensionality have to be taken into consideration:

- Feature redundancy leads to bad clustering. The results are worsted especially if the number of clustering data is low and the number of features is large;
- Irrelevant features also may introduce noise, which leads to obstruction and degradation of obtained data classification or clustering;
- To discover quality patterns, most algorithms require larger training data set on high-dimensional data set. However, in many data mining applications, the training data is usually rather small.

The feature selection can be extremely useful in reducing the dimensionality of the data to be processed by the clustering algorithms. Feature selection is the process of selecting of a subset of the most relevant features in respect to the task to be performed. It can be also defined as the process of removing of the features redundant or not relevant to the task. Thus, it might bring considerable performance benefits regarding computational time and resources in the subsequent clustering or classification.

Feature selection algorithms divide in three main categories: wrappers, filters and embedded methods [3].

Wrapper methods use a predictive model to score feature subsets. Each new subset is used to train a model. The model formatted is tested against a test set and the evaluation and score it toward its error.

Filters methods use an evaluation function to score a feature subset instead of the error rate. Common measures include the Mutual Information [6], Correlation Coefficient, and Inter and/or Intra class distance.

Embedded methods are a group of techniques which perform feature selection as a part of the model construction process itself.

In the framework of the present study, Filter methods are more appropriate for such exploration work, as they rely solely on properties of the data and do not tune to a specific type of predictive model or algorithm. Two Filter methods for feature selection have been explored for the considered biosignal data – Forward greedy attribute subset selection [4] and Gain Ratio Feature Evaluation [5].

Forward greedy attribute subset selection is a kind of group of greedy hill climbing algorithms. These methods iteratively evaluate a candidate subset of features, then modify the subset and evaluates if the new subset is an improvement over the old. Evaluation of the subsets requires a scoring metric that grades a subset of features. In the study we use Pearson product-moment correlation coefficient for distance measure.

Gain Ratio Feature Evaluation is one of the methods based on entropy metric. It estimates feature weights by examining the data and determines for each feature how much information it contributes to the knowledge of the classes of the training data items.

3.2 Cluster Analysis

Cluster analysis is the process of division of data into groups of similar objects. Generally two different philosophies could be implemented in the task solution – hierarchical group separation and group partitioning [11]:

- Hierarchical Clustering. These methods build a hierarchical tree of clusters, and thus bearing different level of granularity. They are further sub-divided into agglomerative (bottom-up) merging small clusters into larger ones and divisive (top-down) splitting large clusters;
- Partitioning algorithms separate data into several subsets. Partitioning techniques might be additionally divided as:

- Probabilistic clustering. Data is considered to be a model of several probability distributions. Expectation-Maximization (EM) method is well known representative of Probabilistic clustering;
- Objective function optimization. KMeans is well known member, with lots of variation – for example X-Mean and Farthest First;
- Density based. They attempt to discover connected sets of data with similar density, separated by less dense regions. DBSCAN algorithm is a well known example of Density-based connectivity category.

Which of the existing methods to use is a nontrivial question. The choice depends on the considered problem and existing knowledge. Therefore way as a part we investigate different clustering methods according to their ability to deal with EEG data.

4 Brain Data Clustering

4.1 Datasets

In the research, we examine 22 datasets that are subsets of the all originally collected ones. Twenty one EEG datasets corresponds to each one of the 21 EEG Channels. A particular EEG Channel Dataset comprise of 52 EEG Signal instances – 26 for positive pictures and 26 for negative pictures.

We use Java Machine Learning Library (Java-ML) [8, 9]. Before applying the methods, each measure, i.e. each line in the dataset, was normalized. The amplitudes are normalized in the interval $[-1, 1]$.

4.2 Clustering Algorithms

Five different clustering algorithms have been studied. They are selected as representative ones within the respective category described in Sect. 3.2:

(a) category *Hierarchical Clusters* – Hierarchical Clustering (HC);
(b) category *Partitioning Probabilistic Clustering* – Expectation Maximization (EM);
(c) category *Partitioning Objective function optimization* – three algorithms:
  - KMeans (KM),
  - XMeans (XM),
  - Farthest First (FF).

The category *Partitioning Density Based* depends on density parameters, rather than number of desired clusters, so not directly applicable in the study.

Each algorithm is configured to produce two clusters.

The two feature selection methods – Forward greedy attribute subset selection and Gain Ratio Feature Evaluation are utilized as preprocessing data adjustment. As some of the selected features by the two methods differ additional experiments by clustering analysis were carried out in order to settle to a feature vector that comprises the best feature group among the features selected by both methods. The preference is given to feature set {3, 7, 10}. The cluster separation in this data space presents best performance results as the covering of the cluster and class data is largest. Detail results of the attributes selected to these data sets are described in [7].

Here, a case with no feature selection steps is also included for experimental investigation in order to reveal benefits of feature selection. Thus, the three feature selection cases have been applied to the five chosen clustering algorithms. The combination forms fifteen experimental scenarios. Below results of all 15 modes are presented and compared.

4.3 Experimental results

Each clustering produces two clusters. We name Cluster_1 the cluster that has more instances of class_1. The other cluster contains predominantly instances of class_2 and we denominated it as Cluster_2.

We measure the accuracy of the two clusters to cover the known two classes. We asses each cluster by percentage instances correctly clustered to the respective associated class. The overall evaluation is average of both cluster assessments. The results of the experiments are shown in Table 1. Values larger than 65% successful clustering are marked in bold.

Several different aspects of the obtained results are considered and analyzed further.

1. Analysis with respect the implemented clustering algorithm.
   - Clustering in category *Partitioning Objective function optimization* show best categorization result. Among these the KMean algorithm presents best performing results. This observation confirms the usability of the methods in wide range of tasks and particularly to the EEG biosignals data mining.
   - Hierarchical Clustering is on the other pole, rendering no practical value with respect to classes discrimination. The result is obtained for the default algorithm settings and does not exclude possible positive results that other method and settings might bring.
   - Partitioning Probabilistic method *expectation maximization* show mean discrimination capability. The nature of date we investigate may explain the mean result, as data does not fit well to a model of the explored probability distributions.
2. Feature selection analyses.
   - Results confirm the usefulness of the feature selection as preliminary step for the clustering. Results in both feature selection algorithms are notably exceeding the case with no attribute selection.
   - Both feature selection algorithms are fairly peers. Nevertheless, it appears that Forward greedy attribute subset selection performs slightly better. Explanation for superior results may be interpreted regarding difference in search method for attributes selection. The combinative estimation of the several most successful attribute appears superior in comparison with evaluation of each attribute independently on others.

Table 1. Clustering results (accuracy in %) with respect to three feature selection methods and five clustering methods

Attribute selection	None					Forward Greedy					Gain Ratio				
Cluster algorithm	HC	EM	KM	XM	FF	HC	EM	KM	XM	FF	HC	EM	KM	XM	FF
Channel 1	51	61	63	61	51	51	51	63	**67**	51	51	57	61	61	61
Channel 2	51	49	51	51	51	51	57	57	57	51	51	49	55	49	51
Channel 3	51	53	55	55	51	49	55	53	53	49	51	59	51	57	51
Channel 4	49	55	53	53	49	51	53	51	55	55	51	55	55	55	59
Channel 5	51	51	51	51	51	51	61	59	59	53	51	51	49	51	53
Channel 6	51	49	59	53	53	51	59	**65**	61	51	51	59	55	53	51
Channel 7	51	53	61	53	59	51	51	55	57	51	51	57	57	49	61
Channel 8	51	53	53	49	51	51	51	53	49	51	51	53	53	49	53
Channel 9	51	49	49	49	53	51	61	63	63	55	51	51	49	49	59
Channel 10	51	51	53	53	53	51	63	51	51	49	53	63	61	63	51
Channel 11	51	49	59	49	51	51	**69**	**72**	**71**	57	51	**65**	**66**	**65**	49
Channel 12	51	51	51	51	51	53	57	57	57	61	51	53	55	55	53
Channel 13	51	51	51	51	53	51	63	**68**	**67**	55	51	49	51	51	51
Channel 14	51	53	55	53	53	51	57	63	63	61	51	57	49	57	51
Channel 15	51	49	49	49	53	51	55	55	57	59	51	51	53	53	51
Channel 16	53	53	53	49	53	53	57	**73**	59	**67**	53	53	61	57	**65**
Channel 17	51	53	51	53	51	51	53	**65**	**65**	55	51	53	**65**	**65**	55
Channel 18	51	51	53	53	49	53	53	53	53	53	51	51	51	51	49
Channel 19	53	55	53	57	55	51	51	59	53	57	53	55	55	51	55
Channel 20	53	63	61	61	**65**	53	64	**67**	**67**	**69**	53	**67**	**78**	**78**	61
Channel 21	51	49	51	51	49	51	49	53	59	49	49	49	53	53	57

HC Hierarchical Clustering
EM Expectation Maximization
KM KMeans
XM XMeans
FF Farthest First

3. Summary of the channels performing the best discrimination clustering results:
   - EEG Channels 11 (Cz Central) and 20 (Oz Occipital) show the best accuracy (above 65%).
   - Channels 16 (Pz Parietal) and 17 (P4 Parietal) also perform relatively well.

   A plausible neurological interpretation of these results is that the centrally located channels (z direction) appear to be more discriminative in terms of human basic binary (±) emotions across various subjects. Second conclusion is that the parietal channels (P) apparently carry more content for clustering the same human emotional states also across different subjects.
4. Cross-analyses.

- We examine the combination of channel, clustering algorithm and attribute selection method according to the clustering score:
- The best clustering result performs the combination ⟨Channel 20 – KMean/XMeans – Gain Ratio⟩;
- Channel 20 has strong discrimination capability, as revealed above. *Objective function optimization* clusters were also brought out as superior methods. Gain Ratio is implicated in the best result, despite the overall better performance of Forward Greedy method.
- Good clustering for combinations ⟨Channel 16 – KMeans – Forward Greedy⟩ and ⟨Channel 11 – KMeans – Forward Greedy⟩, which could be good suggestion for successful combination of channels and cluster algorithms with respective attribute selection.

5 Conclusions

This paper presents a proof of concept work for the viability to discriminate human emotions by clustering techniques applied on EEG Visually Evoked Potentials (VEP) with or without a preliminary feature selection phase.

The first conclusion is that the addition of the feature selection step has significantly improved the performance accuracy of the clustering algorithms. KMeans clustering clearly outperforms the other technique particularly when combined with Forward Greedy feature selection procedure.

The conclusion that the centrally located channels (z direction) are more suitable to discriminate human emotions with positive and negative valence across several subject are supported by the assumption that these channels (Cz, Pz, Oz) are less affected by noise and perturbations.

Acknowledgements

The work was partially supported by the Scientific Research Fund of the Sofia University “St. Kliment Ohridski” under Contract No. 107/19.04.2013 and the European Social Fund through the Human Resource Development Operational Programme under contract BG051PO001-3.3.06-0052 (2012/2014).

We would like to express thanks to the PsyLab from the University of Aveiro, and particularly to Dr. Isabel Santos, for providing the data sets.

References

1. Berkhin, P.: Survey of clustering data mining techniques. Technical report, Accrue Software, San Jose, CA (2002)
2. Grira, N., Crucianu, M., Boujemaa, N.: Unsupervised and semi-supervised clustering: a brief survey. In: A Review of Machine Learning Techniques for Processing Multimedia Content. Rapport du Réseau d’Excellence MUSCLE (6e PCRD) (2004) 11 pp

3. Talavera, L.: An Evaluation of Filter and Wrapper Methods for Feature Selection in Categorical Clustering. In: IDA (2005) 440–451
4. Caruana, R., Freitag, D.: Greedy attribute selection. In: Proceedings of 11th International Conference on Machine Learning. Morgan Kaufmann, New Brunswick, New Jersey (1994) 28–36
5. Karegowda, A.G., Manjunath, A.S., Jayaram, M.A.: Comparative study of attribute selection using gain ratio and correlation based feature selection. International Journal of Information Technology and Knowledge Management **2**(2) (2010) 271–277
6. MacKay, D.J.C.: Information Theory, Inference, and Learning Algorithms. Cambridge University Press, Cambridge. (2003) ISBN 0-521-64298-1.
7. Georgieva, O., Milanov, S., Georgieva, P.: Cluster Analysis for EEG Biosignal Discrimination. In: IEEE International Symposium on INnovations in Intelligent SysTems and Applications INISTA, Albena, Bulgaria (2013)
8. Java Machine Learning Library. `http://java-ml.sourceforge.net/`
9. Waikato Environment for Knowledge Analysis (WEKA). `http://weka.wikispaces.com/`
10. Frantzidis, Ch.A., Bratsas, Ch., Klados, M.A., Konstantinidis, E., Lithari, Ch.D., Vivas, A.B., Papadelis, Ch.L., Kaldoudi, E., Pappas, C., Bamidis, P.D.: On the classification of emotional biosignals evoked while viewing affective pictures: An integrated data-mining-based approach for healthcare applications. IEEE Transactions on Information Technology in Biomedicine **14**(2) (2010) 309–318
11. Fraley, C., Raftery, A.E. How many clusters? Which clustering methods? Answers via model-based cluster analysis. Computer Journal **41**(8) (1998) 578–588

Extending a BPMN Engine with Evaluation Metrics for KPIs

Kristiyan Shahinyan[1] and Evgeniy Krastev[2]

[1] ComSoft Ltd., 47, Knyaginya Maria-Luiza Blvd., Fl. 1, 1202 Sofia, Bulgaria
k.shahinian@comsoft.bg
[2] Faculty of Mathematics and Informatics, Sofia University "St. Kliment Ohridski",
5, James Bourchier Blvd., 1164 Sofia, Bulgaria
eck@fmi.uni-sofia.bg

Abstract. The Business Process Modeling Notation (BPMN) is a standard for developing process models with visual information about how details of a given operation fit together. Unlike alternative approaches to investigate business processes in academic environment with Event-driven Process Chain (EPC) diagrams, this paper highlights the benefits in using a BPMN process model as a template for a set of acceptable process executions on a particular BPMN engine. The coupling of a BPMN diagram with a BPMN engine yields flexibility in backend system integration and scripting. This paper presents an approach for extending the BPMN engine Activiti with scripting that allows to evaluate the performance of academic organizations in terms of Key Performance Indicators (KPIs) and respective evaluation metrics. The KPIs in an academic environment have been investigated and a case study of a typical business process has been considered. The BPMN model and specific KPIs with appropriate evaluation metrics are being introduced. Further on, by means of scripting the BPMN execution engine is being extended to allow gathering simulation data and analyzing the evaluation metrics in relation to the KPIs.

Keywords: business process, business process modeling, business process execution, simulation, BPMN, Activiti, KPI, evaluation metrics

1 Introduction

Today, improving the performance and the competitiveness of a business organization is closely related to measurement and analysis of a set of Key Performance Indicators (KPIs). The KPIs represent a set of measures corresponding to the organization goals and serve to evaluate critical performance aspects of the current and the future success of the organization [1]. The selection of an appropriate set of values for evaluating of business activities is often related to business organizations in marketing, manufacturing, IT operations and project execution, supply chain management, etc. The same problem concerns academic institutions, as well, where its solution is complicated due to the multi-disciplinary and complex nature of the educational process. Therefore, there are few investigations of this problem. Some of them measure the performance of non-academic support units at higher education institutions [2], while other [3–5] employ KPIs as assessment criteria for concurrent accreditation of academic institutions. It is a common mistake in this case to refer to KPIs, while actually considering Key Result Indicators as Student Satisfaction Rate, Graduate Employment Rate, Overall Quality of Services/Facilities/Resources and similar to them. Unlike KPIs these indicators are the result of many activities executed over a long period of time and provide no information about how to improve these results.

Once KPIs have been correctly defined they have to be monitored and measured on a daily basis as part of the execution of a business process. Conclusions can be made from the thus collected data about how to improve the performance of the organization. It requires looking into the sequence of business operations comprising a business process and developing a business process model [6]. The Business Process Modeling Notation (BPMN) is a widely accepted standard for drawing business process models [7]. A BPMN diagram allows binding KPIs to a particular business operation or a sequence of business operations and provides better understanding of the KPIs data source on behalf of the business people. On the other side, the values of the KPIs can be generated during the process model execution and at this stage developers are most often involved. A process execution engine like Activiti [8] allows bridging the gap between business people and developers. Unlike existing approaches [9–12] that predominantly rely on EPC models in related research problems we use a BPMN model for business process simulation in monitoring and measuring KPIs. Consequently, this approach allows the analysis of the thus obtained data to be used for process improvement and business activity monitoring [14].

The goals of this research are to summarize our findings in modeling and analyzing typical business processes at Sofia University. For this purpose we will introduce a realistic BPMN model of a business process that allows us to introduce KPIs in terms of a set of evaluation metrics in the context of the objectives of the university. Further on we extend the process execution engine of Activiti in order to execute that business process and collect simulation data about the evaluation metrics introduced in the BPMN model. Finally, we discuss the thus obtained simulation data in relation to KPIs of the university.

2 Problem Statement

KPIs are traditionally developed as part of an institution's strategic planning process in conjunction with goals, which are conceived on an institutional level and refined by respective departments. KPIs are related to the academic institution objectives and depend on the level of human interaction and communication with modern information systems, management of resources and especially on the level of automated guidance in performing the key administrative and educational processes. In a similar way, academic institutions like universities regularly measure certain Result Indicators in order to maintain high KPIs in assessment procedures serving usually as a basis for distributing government funds in support for education in different areas [3–5]. In this case academic organizations, similarly to other organizations, set particular organizational goals. In order to manage and improve their performance academic organizations need indicators to measure progress regarding these goals.

KPIs reflect various components of the institution's strategic goals and essentially are expressed in terms of performance indicators monitored on a daily basis. The analysis of the performance indicators usually involves some evaluation metrics for the interpretation of indicator values and constitutes the foundation for making decisions about what to should be done in order to improve performance. Performance of an academic organization can be described [3] in terms of KPIs for program and service performance, student- focused performance, operational performance, financial and market oriented performance.

In this paper we focus on evaluation metrics related to measurements of operational performance. Business processes that affect operational performance may be discovered at all organizational levels starting from the individual level and ending with the senior leader level. Examples of KPIs for operational performance include productivity, waste reduction of resources, workforce turnover. The evaluation metrics for these KPIs includes an indicator referred to as *cycle time* [1, 3, 14]. This indicator evaluates the time spent on completing an assigned activity. The decrease in *cycle time* improves all the KPIs for operational performance and therefore it is among the most frequently monitored indicators. Once a BPMN business model is defined then it should be analyzed or subjected to simulation for discovery of t activities that add value to *cycle time*. Generally, the sources for large values of this indicator are manually executed tasks, redundant storage and retrieval of information, repetition of multiple sequences of activities or activities without upper time limits. In order to discover such sources of large values of *cycle time* it should be possible to introduce evaluation metrics as part of the BPMN model and the business process engine executing the process.

In this research we present an approach to extend a business process execution engine with evaluation metrics for measuring the *cycle time* necessary to complete assigned activities or activities as they are often referred to. To illustrate this approach we consider a realistic business processes in an academic institution from the perspective of a BPMN model and simulate the business process execution with properly selected evaluation metrics. The analysis of the thus obtained simulation data from the executed BPMN model allows us to identify weaknesses in the investigated business processes. Finally, we discuss further steps in the elaboration of this approach.

3 Case Study

Let us consider a common business process at a university, the process of generating a Summary report for research activities at the university. For clarity, we have documented the sequence of activities and the whole business process as it is performed at Sofia University. This report is generated on a yearly basis and summarizes all the research achievements, publications, completed projects, presentations on scientific conferences and so on, of all the lecturers and researchers at the university. Some of the faculties use an Information system, named *The Authors*, which stores this kind of data and reuses it in generating reports similar to the Summary report for research activities at the university. Other faculties collect and process manually this data to prepare such a report. A high-level BPMN diagram of the business process is depicted in Fig. 1. The Summary report for research activities is compiled by the Department for Scientific and Applied Activities in the headquarters of the university. The business process starts by a request issued from this department to all the faculties at the university. Next, each one of the faculties summarizes the data received from the lecturers employed in its departments. Finally, the summary reports from all the faculties are being forwarded to the Department for Scientific and Applied Activities in the headquarters, where the Summary report for research activities at the university is being worked out. Note, that each one of the faculties prepares its Summary report for research activities in parallel with the rest of the faculties. The same refers to all the departments at the university.

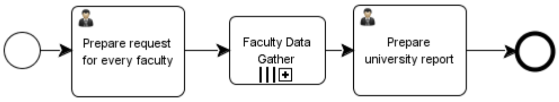


Fig. 1. High level view of the business process for generation the Summary report for research activities

The activity, named *Faculty Data Gather* in Fig. 1, is modeled as a subprocess (defined by the plus mark in the middle) with parallel executing multi-instances (defined by the three vertical lines) of that subprocess. When a request to prepare the Summary report is received at a faculty, then there are two options. If the information system *The Authors* is available, then the report is automatically generated by the information system. Otherwise a request to every department at the respective faculty is being sent. When all departments submit their report, then all the reports get merged into the Summary faculty report.

The *Faculty Data Gather* subprocess expansion is depicted in Fig. 2, where the activity, named *Department Data Gather*, is being introduced. This activity is modeled as a multi-instanced subprocess, executed by every department in every faculty. When a department receives a request, a request is sent to each one of the lecturers. All the lecturer responses are summarized in the department report. The subprocess *Prepare Personal Report* is depicted in Fig. 2 and executed as a multi-instance subprocess, as well.

Now, let us consider the KPIs for Operational performance, namely, Productivity and Workforce turnover with respect to the above described BPMN model. These indicators depend on the *cycle time* performance indicator. It is common

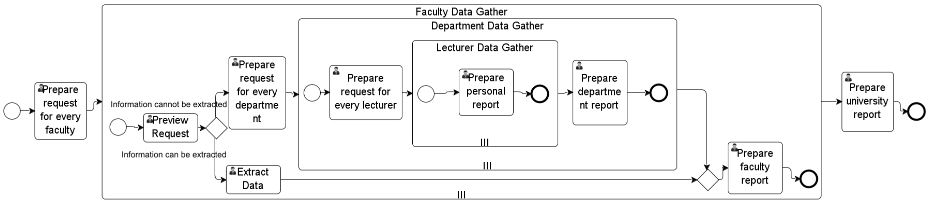

Faculty Data Gather
Department Data Gather
Lecturer Data Gather
Prepare request for every faculty
Information cannot be extracted
Preview Request
Information can be extracted
Prepare request for every departme nt
Extract Data
Prepare request for every lecturer
Prepare personal report
Prepare departme nt report
Prepare faculty report
Prepare university report
III
III
III

Fig. 2. Faculty Data Gather subprocess

to use evaluation metrics in such cases [13]. Similarly, in order to measure this indicator we introduce the following evaluation metrics:

- Total count of an activity execution defines how many times a single activity is executed in a single process execution instance.
- Total time for executing an activity and a subprocess defines how much time is needed for parallel execution of every activity and subprocess in a single instance of the process.
- Average time for executing an activity and a subprocess defines the average time needed for parallel execution of every activity and subprocess.

4 Using a BPMN Execution Engine

There are a lot of BPMN engines, but among them we have selected Activiti [8] because it is open-source, entirely BPMN oriented and there are frequent stable releases. In order to execute a business process with evaluation metrics we have considered the following approaches for simulation of the process execution:

- Extend the Engine's activity execution code. Using this approach the extended code gets *invoked* for every activity that is started by the engine. Currently we are not interested in this option, because our process is really simple and includes only User Activities. This is useful when dealing with different types of activities that should not involve user interaction and the executor is not available (for example using an activity that invokes a Web Service, but this Web Service is not available during simulation).
- Use delegation tasks. Modify the BPMN diagram so that instead of defining what a user should do, we are always invoking our implementation. This is not a good option, because it includes modification of the diagram accepted by the business users and leads to desynchronization of the real-life process diagram and the one used for simulations and testing.
- Act as a user. It involves iteration over the process execution and lookup for unassigned user tasks. If there is any unassigned user task available, it claims it, waits for some time and completes it. Using this approach we are actually acting as a real user. The tricky step here is *waits for some time*, where time can vary between the minimum and the maximum execution time, defined by the business users.

For the purposes of this paper we have chosen the third approach, because:

1. we have our time intervals, defined by business user on the basis of benchmarks;
2. our process contains only user tasks;
3. there is no direct integration with external information systems, which could be potentially unavailable during our simulation process.

Thus, we have defined the following scenario for performing a simulation with Activiti of the business process displayed in Fig. 2:

1. Generate a random number of departments for every faculty (we have fixed 16 faculties), where their number varies between 7 and 16.
2. Generate a random number of lecturers for every department, where their number varies between 4 and 10.
3. Start the process in the engine injecting all the variables required.
4. Loop over the unassigned tasks:
   (a) Claim the task.
   (b) Wait for some time.
   (c) Complete the task.
   (d) If the task is *Prepare university report*, stop iteration (this task is followed by the processes' end event.
5. Extract the following execution information, which one we refer to as evaluation metrics:
   (a) execution count for every element in the diagram;
   (b) average execution time for every activity and subprocess;
   (c) total execution time for every activity and subprocess.
6. Generate a modified BPMN image for every evaluation metric.

The above algorithm allows testing the execution of the business model (Fig. 2) with realistic simulation data and thus obtaining quantitative evaluation metrics data. For instance, to compute the time related evaluation metrics we have used benchmarks for the minimum and the maximum time for manual execution of activities in report generation, coordinated with employees involved in the real-life execution of the process: Additionally we assume that there are 16 faculties, 196 departments (randomly generated between 7 and 16 per faculty), 1387 lecturers (randomly generated between 4 and 10 per department) and a single faculty using an information system for extracting report data.

5 Analysis of the Computed Evaluation Metrics Data

It is important to notice scenario for business process execution treats all of the subprocesses in Fig. 2 as multi-instanced and asynchronous, i.e. all faculties prepare their reports at the same time, and all departments and lecturers are working the same way. Therefore the total estimations cannot be calculated with simple multiplication. Moreover, this timing is for actual work to be done. This is one of the major differences with existing approaches. The advantage of the proposed approach for evaluating metrics of a given KPI is that it provides realistic numeric data. On the other side, this approach allows to bind evaluation metrics to activities that are assumed to contribute to large *cycle time* values and measure their overall effect, for instance, on Productivity and Workforce turnover. In terms of simulation we can establish numerical values for the Best, Worst and Perfect execution of a business process:

- Best case execution, where we assume the shortest time for an activity execution.
- Worst case execution, where we assume the longest time for an activity execution.
- Perfect case execution, where we assume all the faculties are able to prepare by means of *The Authors* external information system.

Further on, the Best and the Worst case simulation results can be used to define precisely the KPI objectives in the execution of the business process in real-life. Accordingly, the Perfect case execution results provide indication for possible business process improvement, which in the here considered case study can be achieved by implementing *The Authors* information system in each one of the faculties of the university.

6 Conclusion

In this paper we have presented an approach for extending a process execution engine for the purpose of simulating an existing academic business process. The analysis of the thus obtained performance indicator values allows measuring the improvement of KPIs for Operational performance. The proposed approach implementation is illustrated by means of a sample case study of a realistic business process in an academic environment. Such processes are not easy to analyze, model, optimize and simulate because they are too complicated and require the engagement of a lot of employees. We have modeled this process using BPMN 2.0 and defined some KPIs and evaluation metrics for Operational performance and quality. Using this approach we are able to go further and use these KPIs and metrics in other processes and compare results and optimization consequences. We have outlined different approaches in business process simulation and described how to use BPMN process execution engines in order to simulate user activities. A *cycle time* performance indicator and related evaluation metrics have been proposed for the purpose of improving KPIs. The here proposed approach provides quantitative metrics for evaluating the Operational performance of a business process in terms of *cycle time* unlike other approaches that use heuristic approaches to evaluate approximately this key process measure.

A future extension of the BPMN standard, where the business process model can integrate the required KPIs and the performance indicators, would be quite useful for the both business people and developers. It would allow real-life diagrams coming directly from business to be simulated directly to a BPMN process execution engine and a BPMN Designer and thus allowing business process improvement in terms of KPIs.

Acknowledgements

This work was supported by the European Social Fund through the Human Resource Development Operational Programme under contract BG051PO001-3.3.06-0052 (2012/2014).

References

1. Parmenter, D.: Key Performance Indicators (KPI): Developing, Implementing, and Using Winning KPIs, 2nd ed. John Wiley (2010)
2. Hanover Research Council: Key Performance Indicators for Administrative Support Units (http://www.hanoverresearch.com) (2010)
3. Baldrige Performance Excellence Program: 2011–2012 Education Criteria for Performance Excellence. National Institute of Standards and Technology (June 2013) (http://www.nist.gov/baldrige/publications/upload/2011_2012_Education_Criteria.pdf)
4. National Evaluation and Accreditation Agency: Criteria System for Institutional Accreditation of Higher Schools (June 2011) (http://www.neaa.government.bg/en)
5. Beatty, A., Koenig, J.A. et al.: Key National Educational Indicators. The National Academies Press (2013) (http://www.nap.edu/catalog.php?record_id=13453)
6. White, S.A., Miers, D.: BPMN – Modeling and Reference Guide. Future Strategies Inc. (2008)
7. Vidovic, D.I., Vuksic, V.B.: Dynamic Business Process Modeling Using ARIS. In: Proc. of the 25th International Conference on Information Technology Interfaces ITI (16–19 June 2003) 607–612
8. Rademakers, T.: Activiti in Action. Manning (2012)
9. Dehnert, J., van der Aalst, W.M.P.: Bridging the gap between business models and workflow specifications. International Journal of Cooperative Information Systems **13**(2) (2004) 289–332
10. Rebuge, Á., Ferreira, D.R.: Business process analysis in healthcare environments: A methodology based on process mining. Information Systems **37**(2) (2012) 99–116
11. van der Aalst, W.M.P., Reijers, H. et al.: Business process mining: An industrial application. Information Systems **32**(5) (2007) 713–732
12. Brander, S., Hinkelmann, K. et al.: Mining of Agile Business Processes. In: Artificial Intelligence for Business Agility, Proc. of the AAAI 2011 Spring Symposium (March 2011) 9–14
13. Mendling, J.: Metrics for Process Models. Springer-Verlag (2008)
14. Friedenstab, J.-P., Janieschy, C. et al.: Extending BPMN for Business Activity Monitoring. In: Proc. of the 45th Hawaii International Conference on System Sciences (2012) 4158–4167

Users and Internet of Things for Wellbeing
Smart Health Cardio Belt

Eugenia Kovatcheva[1], Atanas Georgiev[1], Mirolyuba Madjarova[1],
Roumen Nikolov[2], and Alexander Chikalanov[2]

[1] Faculty of Mathematics and Informatics, Sofia University "St. Kliment Ohridski",
5, James Bourchier Blvd., 1164 Sofia, Bulgaria
{epk,atanas,mira}@fmi.uni-sofia.bg
[2] University of Library Studies and Information Technologies,
119, Tsarigradsko chaussée Blvd., 1784 Sofia, Bulgaria
{r.nikolov,a.chikalanov}@unibit.bg

Abstract. This paper deals with the opportunities that the advanced telemedicine and information technologies can give for attending to the needs of patients with cardiovascular diseases – the leading cause of deaths worldwide. The selected pilot case presents a Smart Health Cardio Belt (SHCB) system developed by the Bulgarian team on the basis of the existing TEMEO prototype. Through the SHCB system, the patient is linked to the medical centre which monitors the patient 24 hours per day while in the same time, the patient is capable in accomplishing all activities of his normal day. The medical centre reacts adequately when a critical event appears. The SHCB system is based on the existing and new developed Internet services. They are piloted, fine-tuned and implemented under the FP7 ELLIOT Project. The paper presents the results from the interviews with final users: patients and medical stuff how helpful is the proposed SHCB. The impact of the Internet of Things is analysed as well as their functionality and effectiveness, exploring the integrity of the social, intellectual, cognitive, economical, legal and ethical aspects of its usage.

Keywords: telemedicine, Internet Of Things (IOT), Smart Cardio Belt

1 Introduction

The Cardiovascular Diseases (CVDs) are globally number one among those causing death: more people die annually from CVDs than from any other disease. In 2008, an estimated 17.3 million people died from CVDs which represents 30% of all global deaths. Each year 9.4 million deaths or 16.5% of all deaths can be attributed to the high blood pressure. This includes 51% of deaths due to strokes and 45% of deaths due to coronary heart diseases [1].

The problem for prevention against CVDs is a hot topic nowadays. During the day everyone has diversity of activities and emotions which can influence the heart itself and the entire cardiovascular system. Its continuous monitoring

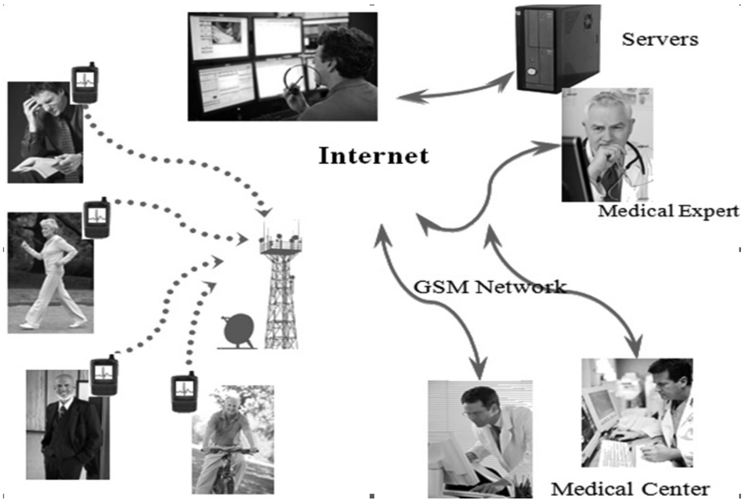


Fig. 1. Usage of Smart Health Cardio Belt

could prevent the negative events. If the patient is linked to an observing medical centre, the specialists there can react when indications for critical events appear (Fig. 1).

These were the main considerations which gave the impetus for development of the Bulgarian Smart Health Cardio Belt (SHCB). This pilot is developed and improved under the implementation of the FP7 ELLIOT (Experiential Living Labs for the Internet of Things) Project [2].

2 Background

The technology for monitoring the cardiovascular system is not new. The cardiac *Holter* sensor exists since early 1960s. But it worked offline, the patient data were collected in it and Medical Doctor had access to them after certain amount of time.

The first Bulgarian prototype of the cardio belt monitored online 24 hours per day was developed three years ago by TEMEO Company. This prototype became the basis for further development of the SHCB system.

The SHCB pilot had been selected after thorough analyses considering the public needs not only in Bulgaria but in Europe wide, the level of the necessary technologies' development and the possible impact of the product for overcoming the current gaps and lacks of the existing TEMEO prototype.

The entire applied research process was actively supported by the end-users who made substantial contribution to a number of improvements. Attracting them to participate in this collaborative work, the end users were split in several categories based on their different consumer interests (medical staff, doctors, patients, hospital administrators).

During the second round of the Smart Health Cardio Belt (SHCB) system experimentation, interviews with the patients and medical staff had been carried out. The purpose of the interview was to gather information on the usability of the new sensors' functions and the compatibility of the data with the ELLIOT platform [2].

3 Pilot Structure, Data Pre-processing and Data Analysis

The interrelations between the components of the system and its functionality are presented in Fig. 2. The patient is equipped with cardio-belt Sensor and Android device. The raw data transmitted from the cardio-belt sensor and Android device to the Service IIS is collected in a cloud.

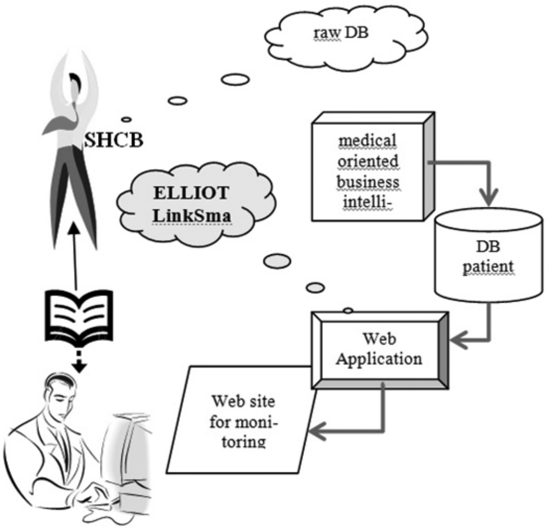


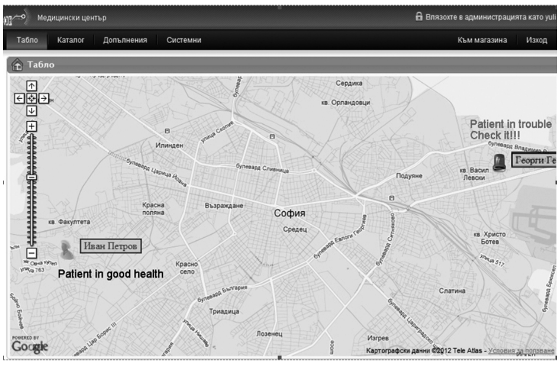
Fig. 2. Functionality of SHCB

The sensor SHCB transfers the data each five minutes, helping to define the initial diagnostics for:

- Absolute arrhythmia (atrial fibrillation);
- Missed heartbeats;
- Bradycardia;
- Tachycardia;
- Atrial fibrillation.

and the corresponding ECG graphic is drawn and provided to the doctors.

An Android device indicates the patient's global position (Fig. 3). In case of emergency it sequentially dials to three phone numbers. For the improvement



Patient in trouble
Check it!!!

Patient in good health

Fig. 3. The patient's position and status

of the Android device design and functions, the patients as end-users are also involved following the methodology of the Living Lab. The initial idea was that when a patient is in a trouble, (s)he should press *a red button for help*. Later on, more useful solution was found – simply to leave the device on the floor. The same effect would be achieved if the patient is not able to react. In case of an accident, a message *Are you OK?* appears on the device's screen. If the patient is able to reply on the message – (s)he is in a relatively stable condition. Otherwise, the device starts dialling to the Medical Centre, sends a message to the Doctor's screen and shows *a red lamp* on the map in order to catch medical staff attention.

3.1 Graphical User Interface

In the first phase of the pilot, three groups of users were identified as potential users of the web-platform: Hospital Administrator, Medical Centre and Medical Doctor. The next step is to visualise the patient's environment. The end-users can read the diagnoses and the relevant treatment without detailed view on the observed data.

Each user has a username and a password in order to have access to the hospital's platform.

The hospital administrator can:

- add patients, doctors and devices;
- arrange an appointment of the patient to a doctor;
- manage doctors' appointments.

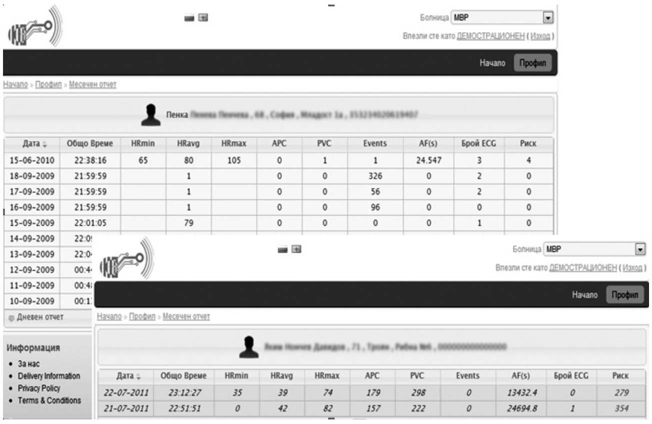
The *doctors* can see only their patients and can read the corresponding data from the devices, which helps them to put right diagnose and to prescribe treatment.

The main goal of the *Medical Centre* is to monitor the observed patients and to react (e.g. call the doctor) if and when necessary.

So far, the collected data by the sensor and the Android device is readable only and used for further diagnostics of absolute arrhythmia, missed heartbeats, bradycardia, tachycardia and atrial fibrillation.

3.2 Data from the Cardio Belt

The data from the belt are structured in monthly reports (Fig. 4) where an average value (per day) is shown, as well as daily reports on each five minutes (Fig. 5).



Болница МВР
Влезли сте като ДЕМОСТРАЦИОНЕН (Изход)
Начало
Профил
Начало » Профил » Месечен отчет

Пенка [redacted]

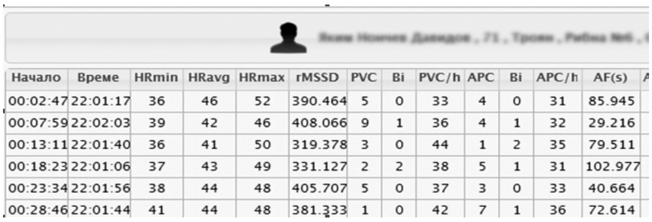
Дата	Общо Време	HRmin	HRavg	HRmax	APC	PVC	Events	AF(s)	Брой ECG	Риск
15-06-2010	22:38:16	65	80	105	0	1	1	24.547	3	4
18-09-2009	21:59:59		1		0	0	326	0	2	0
17-09-2009	21:59:59		1		0	0	56	0	2	0
16-09-2009	21:59:59		1		0	0	96	0	0	0
15-09-2009	22:01:05		79		0	0	0	0	1	0
14-09-2009	22:0									
13-09-2009	22:0									
12-09-2009	00:4									
11-09-2009	00:4									
10-09-2009	00:1									

Дневен отчет
Информация
За нас
Delivery Information
Privacy Policy
Terms & Conditions
Болница МВР
Влезли сте като ДЕМОСТРАЦИОНЕН (Изход)
Начало
Профил
Начало » Профил » Месечен отчет

[redacted]

Дата	Общо Време	HRmin	HRavg	HRmax	APC	PVC	Events	AF(s)	Брой ECG	Риск
22-07-2011	23:12:27	35	39	74	179	298	0	13432.4	0	279
21-07-2011	22:51:51	0	42	82	157	222	0	24694.8	1	354

Fig. 4. Monthly report



[redacted]

Начало	Време	HRmin	HRavg	HRmax	rMSSD	PVC	Bi	PVC/h	APC	Bi	APC/h	AF(s)	A
00:02:47	22:01:17	36	46	52	390.464	5	0	33	4	0	31	85.945	
00:07:59	22:02:03	39	42	46	408.066	9	1	36	4	1	32	29.216	
00:13:11	22:01:40	36	41	50	319.378	3	0	44	1	2	35	79.511	
00:18:23	22:01:06	37	43	49	331.127	2	2	38	5	1	31	102.977	
00:23:34	22:01:56	38	44	48	405.707	5	0	37	3	0	33	40.664	
00:28:46	22:01:44	41	44	48	381.333	1	0	42	7	1	36	72.614	

Fig. 5. Daily report

The data indicating that there might be certain problem with the patient appears in red colour in the relevant rows of the table.

For one epoch (five minutes – first note starting time; ending time) the system reports patient's data based on the collected data by the sensor. They are very important for identification of patients' conditions. Some of these data are:

- Hrmin (minimal pulse rate), Hravg (average pulse rate), Hrmax (maximum pulse rate);
- rMSDD – average square between differences in two neighbour RR intervals in [ms];
- APC – number of supraventricular extrasystoles per 24 hours;
- APC/h – number of supraventricular extrasystoles per one hour;
- PVC – number of ventricular extrasystoles per 24 hours;
- PVC/h – number of ventricular extrasystoles per one hour;
- Events – how many times the Android device is fallen down;
- AF(s) – the time from one epoch when the patient was in absolute arrhythmia;
- AFHRmax – maximum pulse for the moments with arrhythmia.

The last column presents the Risk calculation in range from 1 to 5 [3].

The last data from the device is ECG. It is visualized as a real ECG diagram (Fig. 6). There are possibilities to observe an ECG in more details by zooming it.

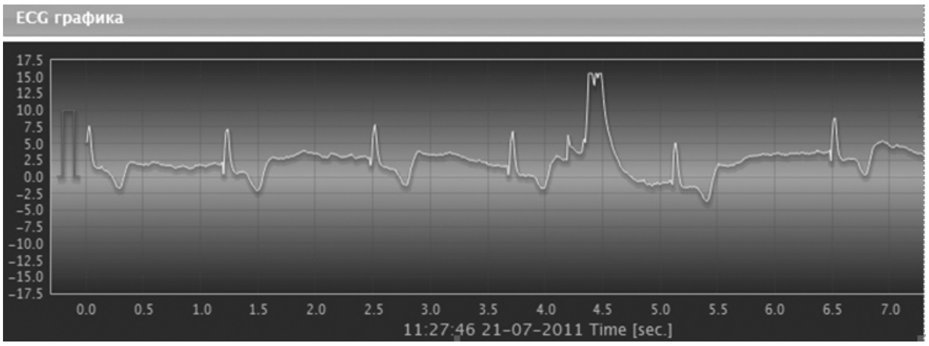


Fig. 6. ECG

3.3 Calculation

The system identifies appearance of the critical events – *true case* – in patient health status (as it was mentioned above, Fig. 3) and analyses its status for a certain period of time according to the formula:

$$\text{IF } 0.5^*(\% \text{ of true cases}) + 0.3^*(\% \text{ of cases with} < 60 \text{ s time for reaction}) + \\ 0.2^*(\% \text{ of cases with} > 10 \text{ patients logged}) > 90\% \text{ THEN GOOD.}$$

The data are normalized and calibrated for this time period. In the SHCB case the appearance of events (for the patients) is not so often (fortunately) and now the formula is applied for a required period of time.

4 Interviews with Patients

Several interviews were carried out for identifying the usefulness of the SHCB by the potential users: patients and medical staff. The interviewed patients have the following age distribution: 40% of them are between 50 and 60 years old; 30% are between 40 and 50 years old; 15% are between 60 and 70 years old; the same (15%) is the percentage for the interviewed between 30 and 40 years old. There were no patients younger than 30 years and older above 70 years.

According to the health status of the interviewed patients, the distribution is as follows: 42% of the interviewed were patients with cardio diseases; 17% were patients with an infarct or bypass; 8% are patients with transplant cardio key-valve; 8% were with other cardio diseases; 8% were patients with other diseases; and 17% were patients in good health status without any disease.

For the entire set of the questions, the interviewed had to present their opinion using the scale from 1 to 5, where 5 is the highest score given to the answer.

4.1 Attitude towards Cardio Belt Usability

As a whole, the interviewed believe in the cardio belt usefulness, because they consider it improves their health condition. There was no answer estimated with 1 (fully negative) and in the same time, there was no answer evaluated with 5.

The majority of the interviewed consider that by using the belt, they obtain better diagnoses.

The patients confirm that the use of the cardio belt contributes to more precise diagnosis (Fig. 7) which leads to the improvement of their health status because they receive more adequate treatment.

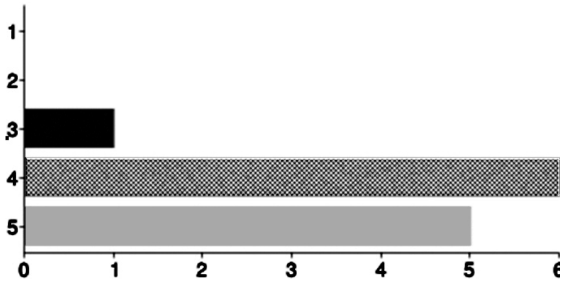


Fig. 7. Attitude towards interrelation between diagnosis (using cardio belt) and health status

4.2 Interviews with Medical Staff

The interviews with the medical staff aimed to receive back information on the use of the Smart Health Cardio Belt system. According to the answers received we can conclude that 60% of the interviewed medical staff uses the SHCB system for observation mainly while 40% of them use the platform for diagnostic activities.

The majority of the interviewed medical staff are fully aware in the usefulness of the SHCB system (Fig. 8).

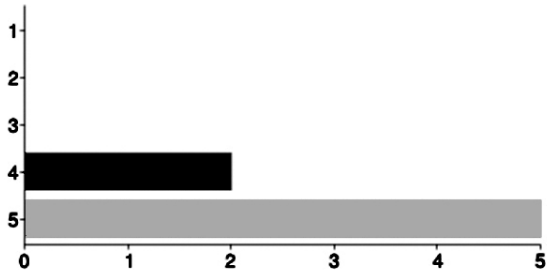


Fig. 8. Attitude towards the level of usability of SHCB system

The new improved platform is highly evaluated by all interviewed.

As a whole, the interviewed give a high score to the system stating that it supports them for the patients who have cardiac problems. The interviewed consider that the SHCB system does not support them with information on the full health status of the observed patient. The medical staff finds it intuitive and easy to use.

5 Results and Inferences

The prototype development of the SHCB was split into three phases. The first one was dedicated to the database creation and connection with a sensor. During the second phase the visual data was maintained and interpreted. The third phase was devoted to the improvement of the patient Android device performance. Some individual interviews with the involved patients and doctors were organised. Their suggestions had been used for further specifications.

The further steps will be focused on the quality assurance of the SHCB and on the implementation of the business intelligence module.

The implementation of the Smart Health Cardio Belt gives a new opportunity for detecting the small aberrations in the behaviour of the patient. This contributes to better observation of the health care principle to prevent from disease, and when the patient is sick, to treat him in the most adequate way. The Smart Health Cardio Belt was designed and the new health cloud was established in the virtual space. The most significant outcome is that 89% of the observed patients feel themselves safer with the SHCB.

References

1. World Health Organization: Cardiovascular diseases (CVDs), Fact Sheet No. 317 Updated (March 2013) Retrieved May 9, 2013, from Media centre: `http://www.who.int/mediacentre/factsheets/fs317/en/`
2. FP7 ELLIOT project: ELLIOT. Retrieved May 9, 2013, from Experiential Living Lab for the Internet of Things: `http://www.elliot-project.eu/`
3. Carola, R., Harley, J.P., Noback, C.R.: Human Anatomy and Physiology. McGraw-Hill Publishing Company, New York (1990)

Research of Success Factors for Start-up Companies

Boyan Yankov

Faculty of Mathematics and Informatics, Sofia University “St. Kliment Ohridski”,
5, James Bourchier Blvd., 1164 Sofia, Bulgaria
boian_iankov@abv.bg

Abstract. A new venture success prediction model is proposed based on an overview and analysis of success prediction models, analysis of the venture creation process, and a qualitative research – interviews with company owners. The success prediction model is extended with measurable variables. A survey to statistically validate the success prediction model is currently in progress with 68 responses by owners and managers of Bulgarian companies. A brief profile of the enterprises and their owners are presented. The available data is analyzed with IBM SPSS Statistics and shows a correlation of the company success and the success prediction model variables.

Keywords: technology entrepreneurship, start-up companies, business processes, business model, new ventures, success prediction, NVP

1 Introduction

The success of a start-up company is defined as the return of investment for all stakeholders [1]. Young companies are valuable for the economy to generate growth, job opportunities and innovations. Success prediction for new ventures is a technique applied to increase the efficiency of the new venture creation process, to avoid the possible failure, to minimize the risks and resources spend and to increase the returns. Unfortunately there are no success prediction models and software tools developed for Bulgarian start-up companies.

A new venture success prediction model for Bulgarian companies would be useful to entrepreneurs, business owners, business incubators, university start-up centres, business consultants, venture capitalists and investors to predict the success probability for the new companies and to identify the possible strengths and weaknesses.

2 Requirements for a New Venture Prediction Model

After an analysis [2] of 42 success prediction models a pattern has been identified. The pattern was introduces by Sandberg [3] in his model from 1986 model as shown in Fig. 1.

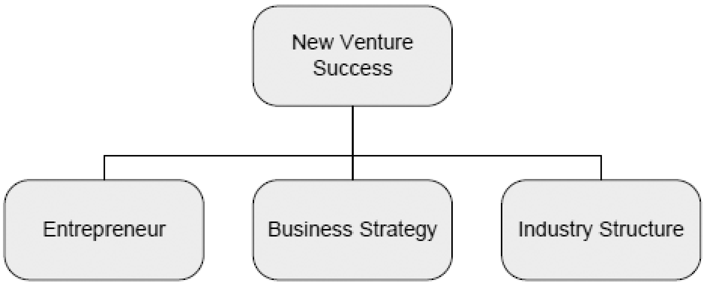


Fig. 1. New venture success model by Sandberg

The model by Sandberg can be illustrated with the formula:

$$\mathrm{NVP} = f(\mathrm{E}, \mathrm{IS}, \mathrm{BS}) \tag{1}$$

where NVP is the new venture performance, E is the entrepreneur, IS is the industry structure and BS is the business strategy. Later studies [4] based on Sandberg include other factors such as the human factor (the entrepreneurial team), the interaction between the company strategy and the industry structure and the available resources.

3 New Venture Success Prediction Model

By analyzing the requirements for a venture prediction model and the venture creation process model [2, 5, 6], an extended new venture success prediction model [7] based on Sandberg [3] is proposed. The model is presented with the formula:

$$\mathrm{NVP} = f(\mathrm{E}, \mathrm{IS}, \mathrm{BS}, \mathrm{R}) \tag{2}$$

where R is a new variable representing the available resources. The other variables are similar to the ones from Sandberg's model: NVP is the new venture performance, E is the entrepreneur, IS is the industry structure and BS is the business strategy. Each of the main categories in the company success prediction model is decomposed into subcategories [7] as it is shown in Fig. 2 – derived by the author.

4 Qualitative Research for the New Venture Success Prediction Model

The new venture success prediction model has been revised with the help of a qualitative research [7] by conducting in-depth interviews with duration of 0:30 to 2:30 hours on a small number of non-representative cases – five owners of innovative Bulgarian companies operating in several industry sectors. The summarized results are presented in Table 1.

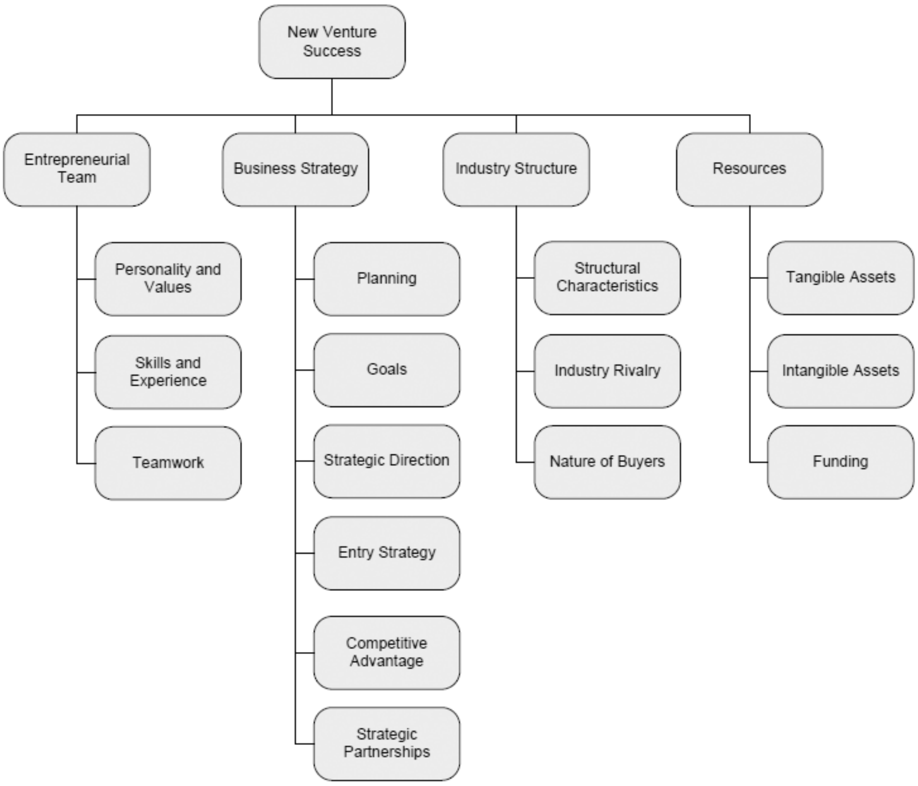


Fig. 2. New venture success prediction model proposed by the author

5 Conclusion and Future Work

The qualitative research confirms that the proposed model for predicting the success of new ventures is accepted by the Bulgarian entrepreneurs as logical and complete. They identified the proposed category "Resources" as an integral and important part of the model. The interviewees expressed a clear interest in the research and a need for a software for predicting the success of young companies and for suggesting possibilities for improvements.

A quantitative research using an online survey (Fig. 3) is conducted to validate the proposed success prediction model for start-up companies. The respondents are owners, directors and managers of start-up and young companies. The goal is to collect responses from at least 200 company representatives during a 6-month timeframe.

According to the results from the quantitative research, the typical Bulgarian entrepreneur is 30–39 aged male with a master degree. 90% of the companies have one to ten employees and are classified as small. The funding source of the companies is usually the founder's own capital.

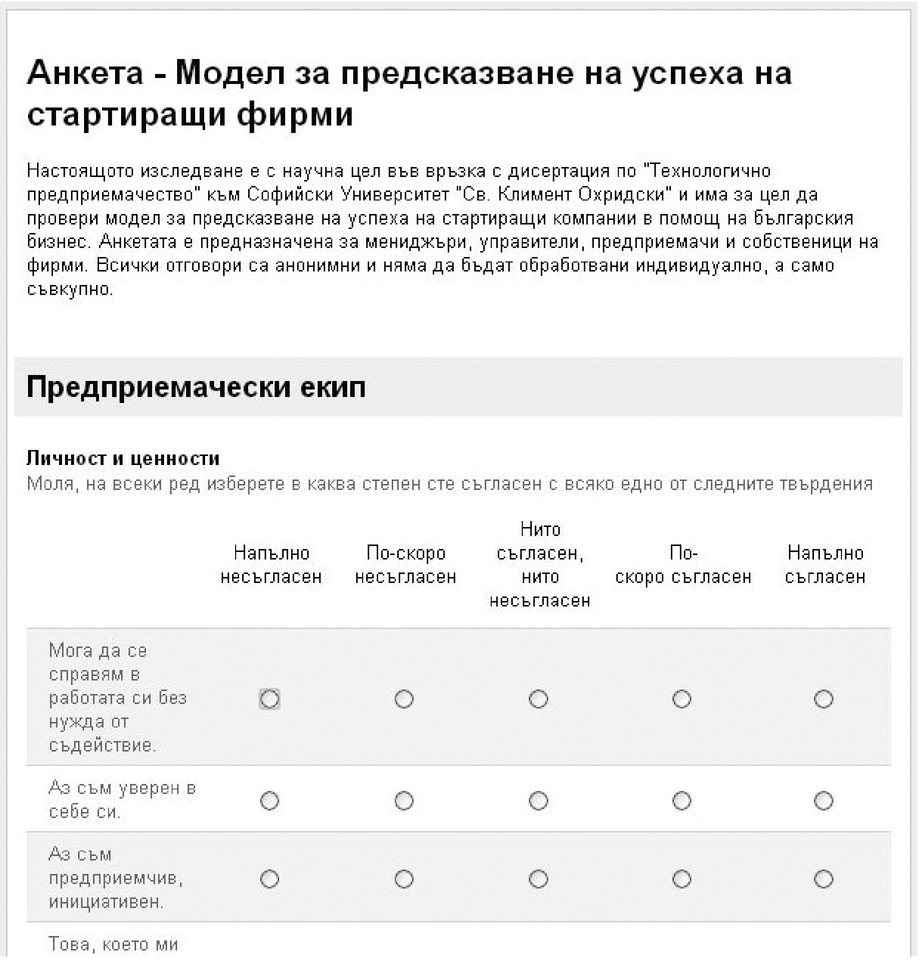
Table 1. Summarized results from the qualitative research of opinions about the proposed NVP model

Qualitative Research Results
The areas of operation of the companies are information technologies, medicine and agriculture.
All respondents are company owners.
The companies were founded 1 to 5 years ago and all of them are currently active.
All respondents described the new venture success prediction model as logical and correct.
The respondents did not find any gaps in the model but they suggested many improvements and additions regarding the individual criteria.
The respondents did not find any unnecessary data in the model.
All respondents think that à software to predict the success of start-ups and young companies would be beneficial. Only some of them are ready to pay for the software but all of them would use it if it was free. They would use the software to help their start-ups and initial researches, for assessment of the company, for localization and improvement of weak points. Some of the interviewees will trust the software but other will first need a proof: information about the algorithm and the underlying logic, details and explanations about the model.
The respondents described most criteria as logical, clear and measurable. They were able to understand and answer most of the questions without difficulties. However they had uncertainties about many questions regarding the team, the personality and the industry. None of them knew their personality type. All interviewees requested definitions for management or marketing terms that they did not understand or misinterpreted. The respondents suggested many improvements and additions regarding the individual criteria.

The current data contains responses by 68 company representatives. An initial analysis of the available data with IBM SPSS Statistics shows correlation of the company success and some of the variables in the new venture success prediction model. This confirms the connection of the new venture success and the variables from the success prediction model. More respondents are necessary to improve the accuracy of the quantitative analysis. Further analysis of the data with statistical software is expected to result in more details and insights about the creation of new ventures.

Acknowledgements

This work was supported by the European Social Fund through the Human Resource Development Operational Programme under contract BG051PO001-3.3.06-0052 (2012/2014).



Анкета - Модел за предсказване на успеха на стартиращи фирми

Настоящото изследване е с научна цел във връзка с дисертация по "Технологично предприемачество" към Софийски Университет "Св. Климент Охридски" и има за цел да провери модел за предсказване на успеха на стартиращи компании в помощ на българския бизнес. Анкетата е предназначена за мениджъри, управители, предприемачи и собственици на фирми. Всички отговори са анонимни и няма да бъдат обработвани индивидуално, а само съвкупно.

Предприемачески екип

Личност и ценности

Моля, на всеки ред изберете в каква степен сте съгласен с всяко едно от следните твърдения

	Напълно несъгласен	По-скоро несъгласен	Нито съгласен, нито несъгласен	По-скоро съгласен	Напълно съгласен
Мога да се справям в работата си без нужда от съдействие.	○	○	○	○	○
Аз съм уверен в себе си.	○	○	○	○	○
Аз съм предприемчив, инициативен.	○	○	○	○	○
Това, което ми					

Fig. 3. Screenshot of a part of the survey for the NVP model

References

1. Bailetti, T.: Technology Entrepreneurship: Overview, Definition, and Distinctive Aspects. Technology Innovation Management Review (February 2012) 5–12 `http://timreview.ca/article/520`
2. Yankov, B.: Overview of Success Prediction Models for New Ventures. In: International Conference Automatics and Informatics'12. ISSN 1313-1850 (2012) 13–16
3. Sandberg, W.R.: New venture performance: The role of strategy and industry structure. Lexington Books, Lexington, MA (1986)
4. Chrisman, J., Bauerschmidt, A., Hofer, C.: The Determinants of New Venture Performance: An Extended Model. Entrepreneurship Theory and Practice **23**(1) (1998) 5–29

5. Carland, J.W., Carland, J.A.: A New Venture Creation Model. Western Carolina University (2000)
6. Abbas, A.A.: An Assessment Methodology for Predicting the Success of Technological Enterprises (2008)
7. Yankov, B.: A Model for Predicting the Success of New Ventures. In: Proc. 5th International Scientific Conference “e-Governance”. ISSN 1313-8774 (2013) 128–135

A New Virtual Private Networks Access Model

Zornitsa Yakova

Faculty of Mathematics and Informatics, Sofia University “St. Kliment Ohridski”,
5, James Bourchier Blvd., 1164 Sofia, Bulgaria
yakova@uni-sofia.bg

Abstract. This paper deals with the concept of a virtual private network (VPN) and the possibility to establish a virtual connection in different than traditional way. Living at a dynamic hi-tech world there is need of a constantly development of new modern scalable and mobile services. The possibility to establish a secure private virtual connection through the network cloud without any dedicated topology and devices or pre-defined setups is a new access model to the private local resources from any external network at any time.

Keywords: virtual private network, vpn, virtual connection, secure access, secure private network, network cloud, new access model, networking

1 Introduction

A common practice is to use a virtual private network (VPN) when you need to utilize the systems and resources – part of your corporate local network from external networks such as the Internet. A VPN network often is a client-server application which handles the secure transfer of data between sites or remote clients via encrypted virtual tunnel. The process is transparent to the participants and communication between them is as they are locally connected.

A VPN is a private network that is created via tunneling over a public network, usually the Internet [1, 2]. Instead of using a dedicated physical connection, a VPN uses virtual connections routed through the Internet from the organization to the remote site. The tunnel is separate logical channel between endpoints and in combination of appropriate protocols supports verification of the identity, integrity and confidentiality achieved by encrypting the traffic through the public network, within the VPN. This type of connection is closed for other network members.

The logical connections can be made at either Layer 2 or Layer 3 of the OSI model [1]. Layer 3 VPNs can be point-to-point site connections or they can establish any-to-any connectivity to many sites.

A VPN is a communications environment in which access is strictly controlled. There are two VPN topologies: a site-to-site which connect entire networks to each other, for example branch office network to corporate headquarters network; and remote-access VPNs which support a client/server architecture where the remote user (VPN client) requires a secure access to the local corporate resources and services via VPN server device at the network edge.

In the site-to-site VPN topology (Fig. 1) the connection devices on both ends (VPN gateways) are pre-configured and the VPN remains static. The process is transparent for the internal hosts. They send/receive normal TCP/IP traffic through their VPN gateways which are responsible for encapsulating and encrypting (decapsulating and decrypting) traffic from one site to another and sending it through a VPN tunnel over the Internet to a peer VPN gateway.

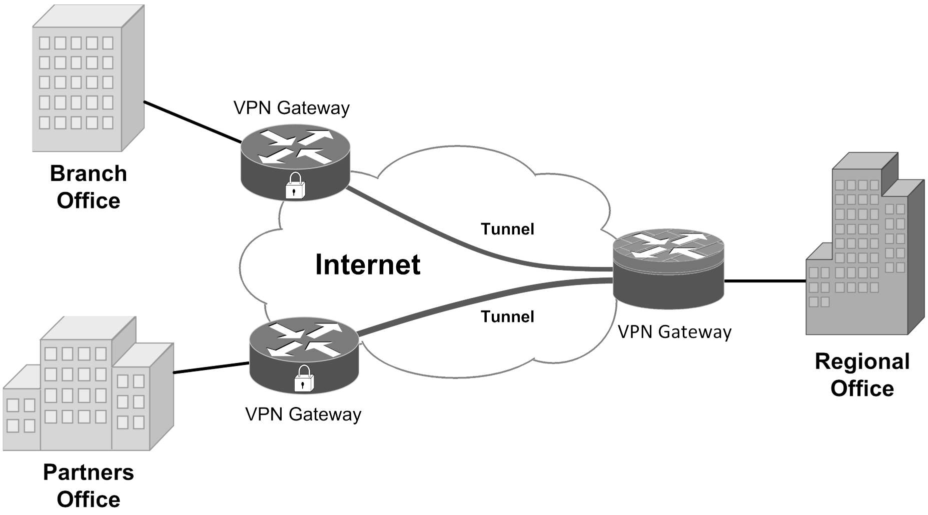


Fig. 1. Site-to-Site VPN [1]

In the remote-access VPN (Fig. 2) configuration setup is assigned dynamically to VPN clients. The VPN connection can be enabled and disabled depending on the telecommuter needs. Permanent connection is not needed all the time. The client is the initiator for the VPN connection and is responsible for establishing the VPN. The VPN server running inside the local network is configured with a virtual network interface on a different subnet. VPN server is waiting for connections on the external network interface where it performs authentication of the VPN client application. If the authentication is successful, VPN client obtains appropriate settings which are part of that virtual network subnet. Then the encrypted tunnel is created between VPN client and VPN server [2]. If client needs other services they could be defined additionally at any time. The change of client's location will not affect the VPN connection.

There are some questions that a traditional approach of accessing the VPN could not answer:

- What will happen if the topology is not known in advance?
- Is it possible to have a VPN access without existence of a configured server?

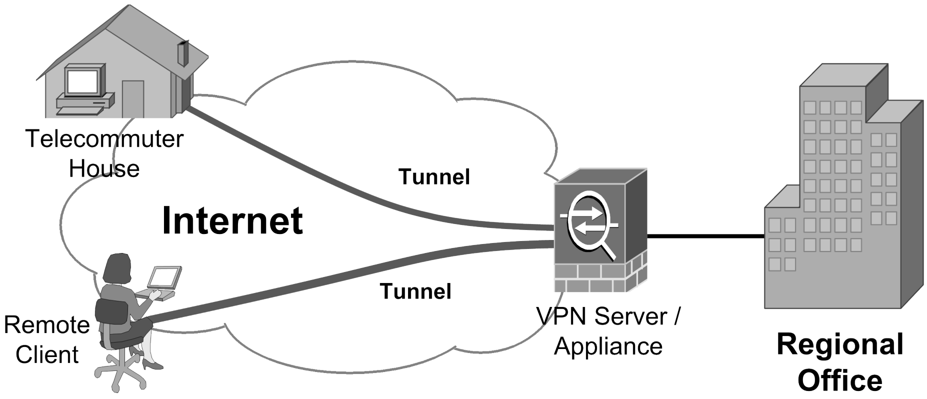


Fig. 2. Remote-access VPN [1]

The aim of current research is to answer the questions exploring the latest innovative information and communication technologies and to find absolutely new possibilities for the design and creation of a new access model to virtual private networks in which the topology is not known in advance and there are no pre-configured VPN servers or VPN gateways.

2 A New VPN Access Model

Nowadays many organizations use VPN so their users can transmit private information over the Internet in a secure way. As a result of a rapid technological development of different complex virtual or cloud-based environments arises the new challenge to access control for VPN [4].

Different varieties of VPN models were made to accomplish the new requirements of the modern networks – to have simple configuration, to be flexible and scalable, to be secure. Such implementation is the proposed on-demand VPN architecture used for communication between multiple VPN users based on a star topology [5]. Another model have realized a web-based interface for full tunneling support VPN architecture, using structured P2P approach for creation and management of the network setups and a central server for all processes related to keys management [6].

The final goal of the proposed model in this paper is the creation of decentralized peer-to-peer shared network infrastructure architecture used securely by users working in collaborative groups at the institutions like university or academia. The participants of such group (who are in the same VPN session) have to communicate directly with each other without a central connection point and the rules they use have to be dynamic depend on situation. In that VPN infrastructure management responsibilities are not just for one entity which improves productivity. The configuration of connection admission control has to be

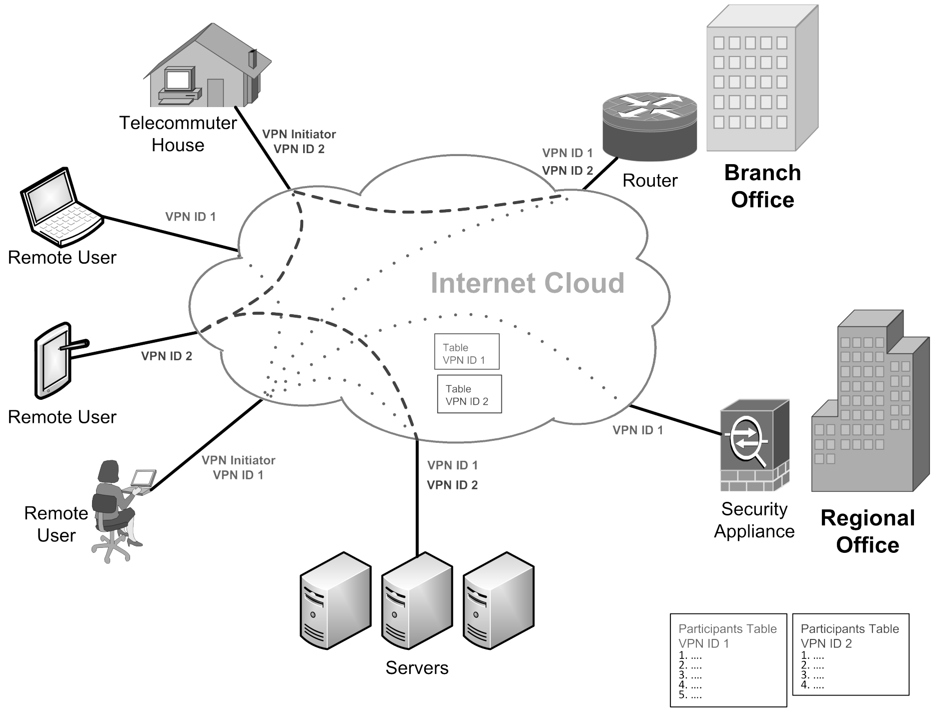


Fig. 3. A new VPN access model

simple and easy to deploy. These features together with an appropriate security setting have to build up a scalable, flexible and cost saving network solution.

The model (Fig. 3) follows the principles of a torrent system. Instead of dedicated server with pre-defined VPN settings and a client who send a request to that server for access to the private channel, there is an initiator who creates a new VPN session itself with a fixed identification number (ID). Peers, part of the same 'network' as initiator, are participants using that VPN session. They should have the same ID like the initiator's. The secure and encrypted connections are setup between the initiators and all remote participants. Tables with all participants associated IDs are created. The information in them is dynamically added and filled out between the participants, synchronized and distributed among all of them via a virtual cloud. In that scenario the topology is not known at the beginning of a process. The main key parts of the model are as follows:

- Creation of a network – the initiator uses a unique string which identifies the network (VPN session ID). A table with all participants using that ID is created and the search for the peers could begin.
- Establishment of the connection to the network – when a participant wants to connect to one VPN session, a new authentication key is generated, similar to SSH public/private key pairs, to establish a secure connection between

the initiator and participant [7]. The key file could be spread out between all participants for that session.
- VPN connection setup – the participants who actually have a role of VPN clients could act as torrent clients who after finding their peers use their own protocol for setup a secure virtual network connection [8]. Security settings of the VPN connection must be agreed to by all participants within the VPN session.

The supposed advantages of the model could be:

- Ease of access – connect to network with ID.
- Decentralization – system without any central coordination.
- Reduce the server and network impact – distributing the traffic load across many computers.
- Flexibility and scalability – the system should function efficiently even with many nodes.
- Failover – not rely exclusively on central servers.

The subjects of a future research are to describe in details the steps for the supposed main components of the model. The questions that have to be answer are as follows:

- How exactly the access to the virtual cloud is accomplished?
- What methods will be used for authentication and encryption? How they will apply to the model?
- What could be the requirements for the implementation and deployment of the model?
- What would be the potential benefits for the user of the technology?

3 Conclusions

Nowadays the new technologies are mostly business and service-oriented [3]. The main developer's purposes are ease of access, simple interface and strong security. Modern networks continue to evolve to keep pace with those changes. Users expect now instant access to company resources from anywhere and at any time. These resources include not only a traditional data traffic but video and voice. There is a need for collaboration technologies that allow real-time sharing of resources and sensitive information between multiple remote individuals as though they were at the same physical location.

The proposed new VPN access model is a step of next generation technologies which will provide the ability for secure and reliable on-demand access to private resources and services in the corporate network through the cloud, no matter location of the user.

Acknowledgements

This work was supported by the European Social Fund through the Human Resource Development Operational Programme under contract BG051PO001-3.3.06-0052 (2012/2014).

References

1. Bollapragada, V., Khalid, M., Wainner, S.: IPSec VPN Design. Cisco Press (2005) ISBN: 1-58705-111-7
2. VPN – Virtual Private Network and OpenVPN. `http://linuxconfig.org/vpn-virtual-private-network-and-openvpn` (August 2013)
3. Rosenbaun, G., Lau, W., Jha, S.: An analysis of virtual private network solutions. In: Proc. of the 28th Annual IEEE International Conference on Local Computer Networks LCN '03, 20–24 Oct. (2003) 395–404. Print ISBN: 0-7695-2037-5, DOI: 10.1109/LCN.2003.1243165
4. Liao, W., Su, S.: A Dynamic VPN Architecture for Private Cloud Computing. In: Proc. of the Fourth IEEE International Conference on Utility and Cloud Computing (UCC), 5–8 Dec. (2011) 409–414. Print ISBN:978-1-4577-2116-8, DOI: 10.1109/UCC.2011.68
5. Koyama, T., Karasawa, S., Kikuchi, Y., et al.: New architecture for a VPN on-demand interconnection system. In: Proc. Of the 8th Asia-Pacific Symposium on Information and Telecommunication Technologies (APSITT), 15–18 June (2010) 1–6. E-ISBN: 978-4-88552-244-4, Print ISBN:978-1-4244-6413-5
6. Wolinsky, D., Abraham, L., Lee, K., et al.: On the Design and Implementation of Structured P2P VPNs. In: Proc. of the Workshop on Service Oriented Computing CoRR. (2010) `http://arxiv.org/pdf/1001.2575v1.pdf` (August 2013)
7. Iyappan, P., Arvind, K., et al.: Pluggable Encryption Algorithm In Secure Shell (SSH) Protocol. In: Proc. of the 2nd International Conference on Emerging Trends in Engineering and Technology (ICETET), 16–18 Dec. (2009) 808–813. E-ISBN: 978-0-7695-3884-6, Print ISBN: 978-1-4244-5250-7, DOI: 10.1109/ICETET.2009.180
8. Wang, G., Cheng, J.: Design of a Semantic Resources Sharing Framework for Structured P2P Networks. In: Proc. of the International Symposium on Computational Intelligence and Design (ISCID), 29–31 Oct. Vol. **2** (2010) 201–204. Print ISBN: 978-1-4244-8094-4, DOI: 10.1109/ISCID.2010.140

Introduction of Enterprise Resource Planning at Bulged

Nikifor Ionkov

Faculty of Mathematics and Informatics, Sofia University "St. Kliment Ohridski",
5, James Bourchier Blvd., 1164 Sofia, Bulgaria
ionkov@fmi.uni-sofia.bg

Abstract. Nowadays, the uncertain environment in Bulgaria faces companies with big challenges to survive on the national market and to be competitive on the single European market. The demands for improving their competitiveness and growth prospects is a reason an increasing number of companies to change their business models and introduce new technologies such as Enterprise Resource Planning (ERP). ERP as an integrated business information systems ensures the overall management of the company resources and business processes, and thus, facilitates its sustainability and competitiveness in a changing market environment. The aim of this article is to present a national case study for solving the problems of a company in the field of printing services. The paper follows a Swiss methodology for case studies writing. Initially, a short presentation of the company is made, followed by the description of the problems faced, and the strategic decision for solving them. Secondly, a concept for designing an ERP system is introduced with brief description of its main functionalities.

Keywords: ERP, concept, business optimisation

1 Introduction

In the knowledge-based economy knowledge and innovation have become essential factors for competitiveness and growth of organizations [1]. As pointed out by Zack [2], knowledge provides advantages to enterprises, which are unique and develop over time within organisational learning, and ensure their stability and prosperity. Therefore, it is not surprising that scholars and practitioners have paid particular attention to knowledge management (KM), and the introduction of various KM techniques and tools in organizations. In particular, Enterprise Resource Planning (ERP) provides an essential tool for managing organizational business processes and avoiding the duplication of knowledge within them. Generally, ERP systems provide a "single solution from a single supplier with integrated functions for major business functions from across the value chain such as production, distribution, sales, finance and human resources management" [3]. ERP facilitates the integration of all internal and external processes of the organization, and thus, contributes for greater efficiency and quality of the services provided to customers. In addition, it ensures better knowledge and information sharing within the organization, and generally contributes for better decision making.

Unfortunately, the high costs of proprietary ERP systems and the need for customization of the products are a reason for many companies in Bulgaria not to use them or just to introduce some basic functionalities [4]. However, faced with the increasing competition in the country and abroad many companies are looking for increasing their efficiency and the overall management of their business processes. The paper presents the case of a Bulgarian company which has a relatively strong market position in the publishing sector, however, faces several challenges due to the lack of appropriate information and communication technologies (ICT) to integrate its business processes, and the information processing and flows within the organization and with its customers and suppliers. The paper initially focuses on the main activities and problems faced presently by Bulged. Second, it presents a concept for developing ERP system for the benefits of the organization. The paper follows a Swiss methodology for case studies writing [1], which was transferred by the authors to Sofia University staff within recent projects. The methodology includes the following phases:

- Company general information, vision and goals, the importance of IT in the company's strategy;
- Launching a project – reasons and problems faced;
- Description of the specific solution, e.g. model, processes, applications and technical implementation;
- Project implementation, including management, software development and maintenance;
- Experience obtained during the implementation, e.g. recognized benefits, degree of achievement of objectives and subsequent amendments;
- Success factors – linked to the specific solution and its practical implementation.

As the project of Bulged is underway, some of the paper focuses only on the initial parts of the methodology, and in the conclusion considers the expected benefits, and the possible success factors.

2 Main Challenges ahead Bulged

2.1 Activities and Business Processes

Bulged was established in November 2001 [5], and offers its customers high quality services in the field of polygraphs – prepress and digital print quality [1]. Bulged works with a wide range of clients, currently their number exceeds 2000, and over 70% of them are long-time customers of the company. The "Bulged" activities include various business processes. The main stages through which the performance of the application on each client are illustrated in Fig. 1. The core business processes of Bulged, like in most companies offering printing services, include:

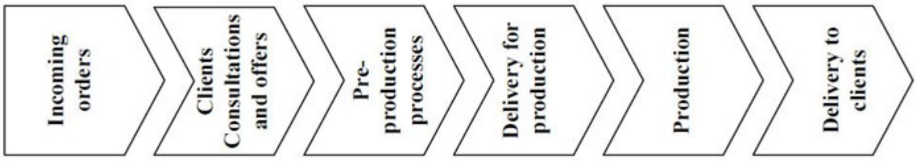


Fig. 1. Bulged business processes

- preparing preliminary services costs – including costs of processes, raw materials, materials and accounting of all costs included in the services;
- determining production norms – defining the resources used for each service;
- tracking of services and processes provided by external contractors/suppliers;
- monitoring of the services implementation in different stages of their delivery to the client;
- accounting and business management;
- preparing activity analysis and statistical information – reports, online real time statistics and graphs for different variables – processed requests, clients served, units load, equipment, etc.
- effective communication and management of customers services – taking orders, communication, sending offers, discounts, etc.

2.2 Problems Faced

The main challenges ahead of Bulged are associated with the dynamics of the sector of printing services and the traditionally strong competition in the industry, as well as the restrictions imposed by the global economic crisis. Therefore, the Bulged leadership stresses the needs for constantly expanding the services range, providing high quality and short deadlines, and maintaining competitive prices. Subsequently, Bulged has invested in up-to-date equipment and technologies, quality management and specialized software solutions for process control. Recently, the company management identified as an emerging need the introduction of advanced software solutions for managing the company business processes, material and financial resources, and the work with customers. A major problem of Bulged is the lack of a system to automate the entire business, and to use only a single software for accounting and stock management. Among its problems could be listed the following:

- unable to prepare a preliminary cost of the services due to the wide range of services (over 150) and the use of more than 45 different production;
- often do not possesses in advance information on cost norms for the services and stock availability;
- unable to automatically trace the suppliers orders;
- lack of possibility of integration of the production software;
- lack of system for managing customer relationships (CRM).

These problems are a major challenge for the successful development of Bulged. In order to overcome them, the company management addresses the need for implementation of an ERR system.

3 ERP System Concept

The main objective of introducing ERR system is to ensure efficient organization of the overall business of the company, as well as efficient management of its resources and work with customers and suppliers, as well as to optimize the production process. The system must have a minimum of systems and databases as shown in Fig. 2. For the automation of work with clients, suppliers and subcontractors will help CRM and Supplier Relationship Management (SRM) systems, while managing stocks, production and finance will be ensured by specialized modules. Integrating these systems would allow a total control of the production process, including:

- traceability of the cycle for execution of client requests and deliveries from subcontractors – from initiation of request/delivery to its completion;
- control and monitoring of production processes and technology sequences, including the design of different types of operations, and specification of their parameters, construction of basic technology plans, and specialized (e.g. printing and finishing), estimation of the necessary materials for implementation of each operation and the entire technology plan expenditure norms, including nomenclatures;

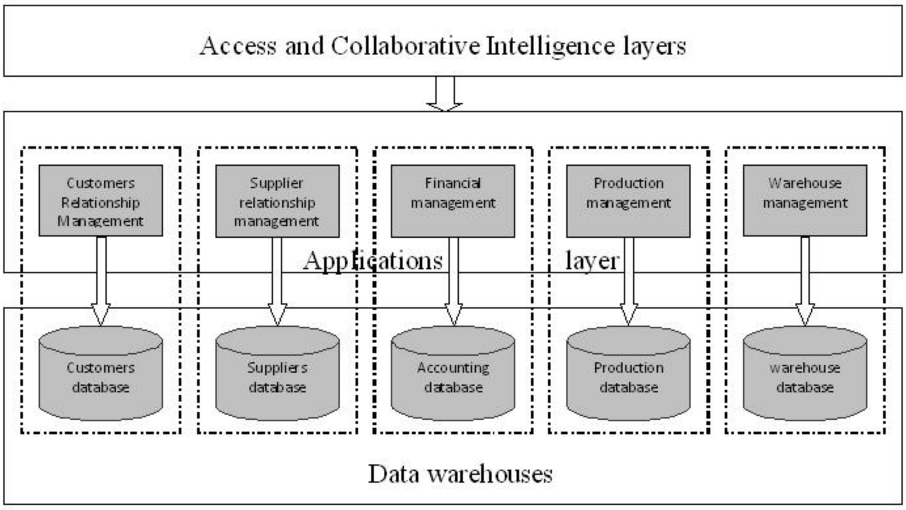


Fig. 2. ERP system concept

- conducting preliminary estimation of the necessary materials, printed forms and instruments for the implementation of each operation, and the time to execute it according to preset standards, as well as the number and type of unit operations, including preparation and completion of the operation. On this basis, it will be possible to calculate the price by taking into account the performance time of the operation and the planned costs of materials;
- providing information on stocks availability, ensuring traceability of materials, semi-finished and finished products, including an entry into the warehouse supply.

The system must be able to generate the following reports and inquiries:

- Report on any technology plan per operations including the amount spent on materials (paper, inks, foils, etc.), the quantity or semi-finished products or ready production, a technology report, on the achieved cost;
- Report of spent materials per orders for a period of time;
- Report of the production schedule for a period of time, clearly showing the excess or saving time for actual production, overspent or saved time for preparation and completion of operations, non-productive time, etc.;
- Report on the work of operators and teams for a specific period of time, including attendance time, overspending or saving production time, and time for preparation and completion of the operation, as well as overspending or saving materials – quantitative and financing;

The presented system concept will ensure all functionalities needed by Bulged. During its design, a full integration in the system of the available software tools will be considered, too. The objective is the new ERP system to monitors the real time implementation, and thus, the marketing unit will be able to provide information to the clients at any stage of the services implementation. On the other hand, after entering into exploitation of the ERP system, serious optimisation of time and resources in the decision making process it is expected, as the system will be able to check stocks in the warehouse, and alarms in case of a lack of a particular commodity. Moreover, the system will raise the efficiency in generating preliminary estimates of the services, prime costs and other information.

The core areas and processes which will be improved by implementing the ERP system include:

- Management of company commercial activities;
- Operations management – due to the ability to generate various reports and inquiries;
- Core business and production processes automation;
- Real-time monitoring and control;
- Quality of printing services, as well as of customer services;
- Customer relations and service;
- Communication between different units and offices;
- Decision-making process related to the company long-term development and investments.

4 Conclusions

The success of the ERP implementation will be ensured by committed leadership, which is in place in Bulged, as well as training of the personnel before the exploitation starts. Regarding technology, it is essential to build on the available tools and ensure smooth transition to the new operations mode. This requires custom design of the software tools and careful planning of data migration and all integration processes. A proper maintenance and service by the IT designer afterwards would be an essential element of the successful implementation.

The ERP concept implementation will ensure vast improvements of management practices, administrative and operational activities of the company, as well as improvement of production – cost ratio. Other benefits will be higher quality of services, optimizing execution time, and reducing prime costs. The company will improve significantly its performance, customers relations and market position, and the ability to consolidate its leadership in the market of high-quality printing services in Bulgaria. The ERP implementation will facilitate Bulged competitiveness for new larger-scale projects. The new investments in the company production and organization in conjunction with qualified and informed staff will be a prerequisite for rapid growth and attracting new customers. As a result, the company will become stronger and more successful competitor in its sector.

Acknowledgements

The author wish to kindly acknowledge the support and information provided by Bulged management for this paper and the financial support ensured by the European Social Fund through the Human Resource Development Operational Programme under contract BG051PO001-3.3.06-0052 (2012/2014).

References

1. Gourova, E., Antonova, A., Nikolov, R. (eds.): Knowledge Management. Bulvest 2000, Sofia, (2012) ISBN 978-954-18-0839-9 (in Bulgarian)
2. Zack, M.H.: Developing a Knowledge Strategy. California Management Review **41**(3) (1999) 125–145
3. Bocij, P., Greasley, A., Hickie, S.: Business Information Systems: Technology Development and Management for the E-Business, 4th ed. Pearson Education Ltd, Harlow (2008)
4. Gourova, E., Antonova, A.: The challenges of e-business development in Bulgaria. In: Proc. of 3rd Balkan Conference in Informatics, 27–29 September 2007, Sofia, Bulgaria, 421–430
5. Bulged, `http://bulged.net/`

Lost in Letters

Temenuzhka Zafirova-Malcheva

Faculty of Mathematics and Informatics, Sofia University “St. Kliment Ohridski”,
5, James Bourchier Blvd., 1164 Sofia, Bulgaria
tzafirova@fmi.uni-sofia.bg

Abstract. The paper presents some aspects of training computer skills with children with cerebral palsy and problems that occur in countries that use the Cyrillic alphabet. The use of different language and alphabet requires mixed computer keyboards. This causes many problems and they are presented in the paper. Discussed is the problem with lack of suitable alternative keyboard devices for students with special needs in countries where the Cyrillic alphabet is used. The solution of this problem by adapting a standard keyboard with labels is also presented.

Keywords: Cyrillic-Latin mixed keyboards, alternative keyboard devices, typewriting problems of children with cerebral palsy, adapt standard keyboard

1 Introduction

Irrespective of the way we communicate – directly or through a number of different technologies, the language – spoken and written – remains the main communication way. Different nations use different languages. According to the research “Ethnologue: Languages of the World” of SIL International in the world today there are 7105 living languages [1, 2]. There are different writing systems used by people. Some languages have been written with a number of different writing systems over the years. There are also writing systems that are used by more than one language. The most widely used writing systems are the Latin, Cyrillic and Arabic alphabets [3].

In the area of computer technology English language has established itself as the dominant language, and so is the Latin alphabet. Design and development of advanced computer technologies comply with users’ national and cultural differences. Accessibility is one of the most important issues in the development of new software and hardware products. Resent versions of operating systems and many software programs offer interfaces in languages other than English, however, the use of English is still essential. The use of languages other than English and alphabets other than Latin still causes some difficulties. These difficulties grow into problems when we talk about users with special needs. This group of users includes people with different kinds of physical, mental or communication disorders. Most of them need assistive technologies to interact with a computer system, but sometimes they are accessible only in English language or Latin alphabet. Uses of standard devices such as keyboards and mice depend on the type and the degree of disability and is strictly individual. Although these problems are unavoidable, the ability to work with a computer is very important to users with special needs. These skills enable to overcome problems associated with mobility, socialization and work. So, skills for computer work are crucial for users with special needs and it is necessary to be acquired in childhood.

In this paper we present some problems that appear with children with cerebral palsy in Bulgaria in the process of their computer education using their native language and respectively Cyrillic alphabet. These problems were observed during the research for training computer skills of children with cerebral palsy. Training was conducted in “Special hospital for prolonged treatment and rehabilitation of children with cerebral palsy – St. Sofia”, Sofia, Bulgaria, for the period of five years with 24 children aged 6 to 16 years.

2 Computer Training of Children with Cerebral Palsy

It is important to clarify what exactly is Cerebral Palsy (CP) and what problems it causes. It is a damage to the developing brain of the newborn, occurring usually before or shortly after the time of birth. CP may affect in varying degrees different areas of the brain, namely the coordination of movement and posture, fine motor skills, speech, intelligence, perception, emotions, sometimes combined with epilepsy, blindness, deafness and other abnormalities. But mainly it affects the person’s ability to move and maintain balance and posture [4, 5].

The main problems in training children with CP to work with computer are caused by muscular weakness, poor motor control of arms and damage of fine motor skills. These problems reflect to children’s ability to control the main input computer devices – the keyboard and the mouse. Problems with intelligence and perception can also affect the ability to work with these devices. But what significance has language and alphabet that children use? People who use non-Latin alphabet are forced to use mixed keyboards (Fig. 1). A typical design of these keyboards is that besides the letters of the Latin alphabet on the keys are printed the letters of the Cyrillic alphabet (for Bulgaria). This kind of organi-

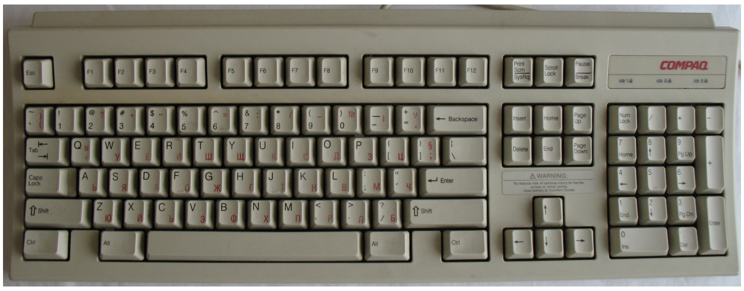


Fig. 1. Mixed keyboard

zation of a keyboard cause many problems with its use in computer training of children with CP.

3 Keyboard and Children with Cerebral Palsy

First of all, using the keyboard requires recognizing the symbols of letters. It also requires reading literacy, which affects the age limit to start training work with keyboard.

Working with a keyboard is very important for children with CP. This is especially true for children with paresis of hands that does not allow them to handwrite, and non-verbal children whose alternative means of communication is by written speech. The use of computers for typewriting and respectively the ability to work with a keyboard are very important for these children. But as mentioned, in Bulgaria only mixed keyboards are used and this causes a number of problems for children with CP.

In general, the design and appearance of the keyboard is quite complex. There are many keys with different symbols. The mixed keyboard becomes even more complicated, because in the symbolic keys of the keyboard on each key there are images of two characters – Latin and Cyrillic. This may cause confusion especially when the characters are not written in different, contrasting colors or font size. There are too many characters on the keys and for children with CP there is a problem with orientation on the keyboard.

Because in the creation of the Cyrillic alphabet some of the symbols of letters are borrowed from the Latin, both alphabets have several letters with the same appearance, but different pronunciation. However, the arrangement of symbols from the two alphabets on the keyboard is different. Latin uses QWERTY layout and Cyrillic uses the Bulgarian standard (BDS) layout. This causes the following problems when children with CP start to learn position of letters on the keyboard:

- two different letters, Latin and Cyrillic, are located on the same key (Fig. 2);
- same letters are located on different keys (Fig. 3).

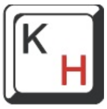


Fig. 2. Two different letters on the same key

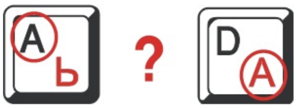


Fig. 3. Same letters on different keys

This is very confusing. Children often wonder which key to choose.

Other specifics of two alphabets are that there are letters whose symbols are mirror images. This adds additional confusion in learning the position of the letters on the keyboard (Fig. 4). Children often confuse the letter "E" with the

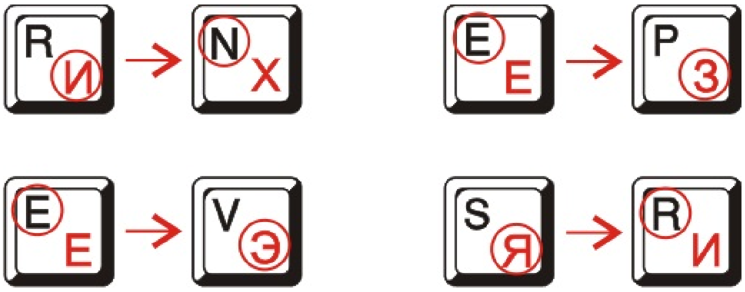


Fig. 4. Mirror images of letters

Cyrillic letter "З" or "Э" (this letter is not used in Bulgarian, it is Russian, but there are historical reasons to be present on the keyboard). Children also replace the Cyrillic letter "И" with the Latin "N" and the Cyrillic "Я" with the Latin "R" [6].

Problems are raised by usage of special keys – Enter, Backspace, Ctrl, Shift, Alt, Caps Lock and ESC. Shortcut keys combinations like Shift and "letter" (for typing capital letter) also are difficult, not only because of the physical motion (pressing two keys at same time), but also because children must remember the key combination and its purpose [6].

In the beginning of our research we highlighted Cyrillic letters on the keyboard by applying red paint on relevant keys. This decision had a major disadvantage – the additional red characters were deleted very quickly and we had to write them again and again.

These problems can be easily overcome by the use of appropriate alternative keyboard devices with large keys, simple design and construction, and printed letters only in Cyrillic.

Alternative keyboards are great, but... Why we could not find such kind of a keyboard?

Alphabetical problem – the Latin alphabet has only 26 letters while the Cyrillic one has 30 letters. Alternative keyboards are designed so that the typing keys are only 26 (Fig. 5). So, for four Cyrillic letters there is no place on such type of keyboard devices. This technological problem makes alternative keyboards unusable for countries with native languages based on Cyrillic alphabet, including Bulgaria.

4 Is there a solution?

When there is still no appropriate alternative keyboard, the only solution is to adapt a standard keyboard. For this purpose, different colored labels with large print of Cyrillic letters are used (Fig. 6). These labels are hard to find in the Bulgarian market, but can be made by order.

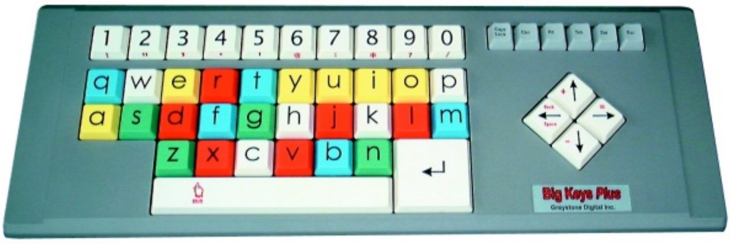


Fig. 5. Special keyboard "BigKeys Plus" – 26 keys for letters

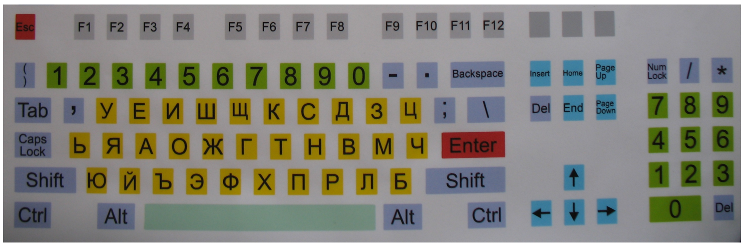

Esc F1 F2 F3 F4 F5 F6 F7 F8 F9 F10 F11 F12
1 2 3 4 5 6 7 8 9 0 - . Backspace Insert Home Page Up Num Lock / *
Tab , У Е И Ш Щ К С Д З Ц ; \ Del End Page Down 7 8 9
Caps Lock Ь Я А О Ж Г Т Н В М Ч Enter 4 5 6
Shift Ю Й Ъ Э Ф Х П Р Л Б Shift 1 2 3
Ctrl Alt Alt Ctrl 0 Del

Fig. 6. Large print colored labels with Cyrillic letters

This solution has more benefits than just hiding the Latin letters from keyboard keys – it overcomes problems arising from the use of mixed keyboard. Labels become a great tool to perform and apply a method for learning position of letters on the keyboard on small groups. During the first stage of adapting the keyboard, labels are used to hide all symbols on the keys. Different colors are used to mark keys with different functions (Fig. 7). Then new groups of letters are added, so that their positions on the keyboard could be easily remembered (Fig. 8). This continues until all symbols which will be used are placed on the keyboard and the adapted keyboard is finally ready (Fig. 8). This process is

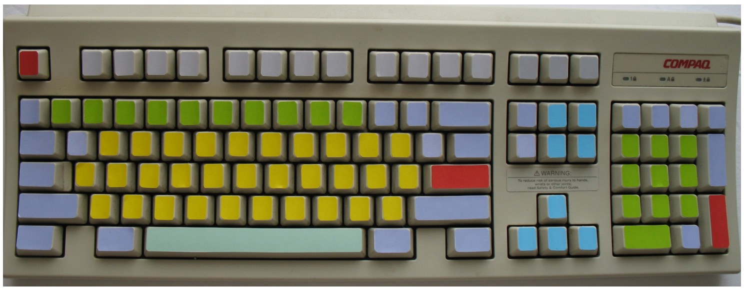


Fig. 7. Distinct blocks of the keyboard are painted with different colors. Keys with same purpose are painted in the same color

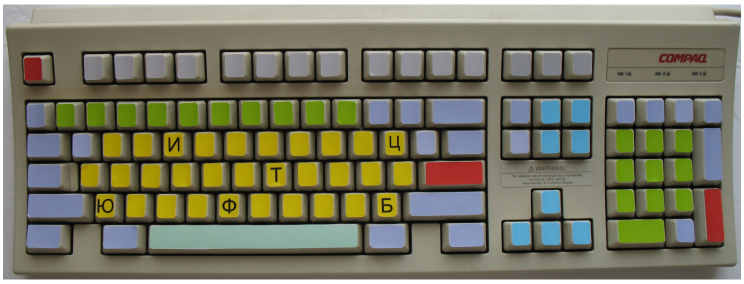


Fig. 8. Learning position of letters in small groups

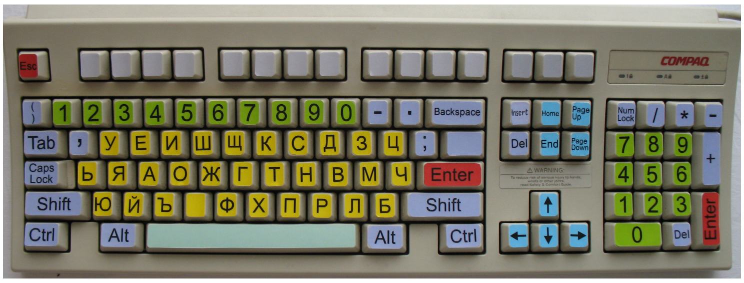


Fig. 9. Adapted standard keyboard

complete when children remember positions of all letters on the keyboard. Then they are ready for the next stage of training – text input.

The following benefits were observed when using labels for modifying the keyboard:

- Using colored labels for contrast, definition of purpose and draw attention:
  - improves visibility by contrast colors for background and letters (in this case black letters on a yellow background);
  - highlights with different colors different parts from the keyboard according to their purpose (letters in yellow, numbers in green, etc.);
  - considers the choice of colors with human perception (yellow – the first noticeable, red – important, etc.).
- Using colored stickers without symbols to facilitate learning positions of letters on the keyboard step by step by adding letters on small groups.

This method of learning position of letters on the keyboard can be combined with appropriate educational software, for example, an educational game that shows on the screen a picture of an object whose name begins with any of the studied letter. As a response the child must enter the corresponding letter from the keyboard. Such games are the games “Letters” and “Words” from the software package “Games House”. This is a special software environment designed and developed during the research for computer education of children with CP mainly to acquire the skills to work with keyboard and mouse [7].

5 Conclusion

Alternative keyboards provide a great advantage in mastering work with a keyboard by children with cerebral palsy, especially when this is the only way to write or it is a major communication tool. While these devices remain inaccessible to certain groups of users, a search for suitable tools for adaptation of standard equipment is the only alternative to overcome the problems of using mixed keyboards in countries with Cyrillic alphabet. The adaptation of standard keyboard with labels appears a very good decision of this problem and it gives the opportunity to use an alternative method for learning position of letters on the keyboard.

Acknowledgements

This work was supported by the European Social Fund through the Human Resource Development Operational Programme under contract BG051PO001-3.3.06-0052 (2012/2014).

References

1. SIL (Summer Institute of Linguistics, Inc), http://www.sil.org/ (27 June 2013)
2. Ethnologue: Languages of the World, http://www.ethnologue.com/ (27 June 2013)
3. Omniglot, Encyclopedia of writing systems and languages, http://www.omniglot.com/ (27 June 2013)
4. Valente, J.A.: Creating a computer-based learning environment for physically handicapped children. Technical Report, Massachusetts Institute of Technology, Cambridge, MA, USA (1983)
5. Centers for Diseases Control and Prevention, Cerebral Palsy, http://www.cdc.gov/ncbddd/cp/index.html (27 June 2013)
6. Ivanov, I., Zafirova-Malcheva, T.: In search of the keys ... In: Digital Tools for Lifelong Learning. Proceedings of the 10th European Logo Conference (EuroLogo'05), Gregorczyk, G., Walat, A., Kranas, W., Borowiecki, M. (eds.), Centre for Informatics and Technology in Education, Warsaw, Poland
7. Zafirova-Malcheva, T.: Development of educational games for special needs education. In: 4th International Conference of Education, Research and Innovation, November 14–16, 2011, Madird, Spain, ICERI2011 Proceedings CD. (2011) 6819–6827 ISBN: 978-84-615-3324-4

Computer Technologies for Accessible Education

Temenuzhka Zafirova-Malcheva and Pavel Boytchev

Faculty of Mathematics and Informatics, Sofia University “St. Kliment Ohridski”,
5, James Bourchier Blvd., 1164 Sofia, Bulgaria
{tzafirova,boytchev}@fmi.uni-sofia.bg

Abstract. The paper discusses the problems with the use of computer technologies in education of people with special educational needs (SEN). It presents different types of software and hardware assistive technologies that improve accessibility to computer system for these users and describes how these tools can be useful in solving educational problems.

Keywords: special educational needs, accessibility, assistive technologies

1 Introduction

Technologies has proved their place in the educational process, not only as an object of study, but also as a tool supporting the educational process in various aspects – management and organization, acquisition of knowledge and skills in different areas, and for different groups of students, including those with special needs – people with communication, emotional and behavioral problems, learning and physical disabilities, and developmental disorders. When talking about students with special needs we usually use the term *special educational needs* (SEN) to denote situations in which it is impossible to apply the standard curriculum. In such cases, we should apply an individual learning approach, consistent with the type of disability and the nature of the learning difficulties.

A few basic components are important when using technologies for education of learners with SEN – the appropriate *hardware* and *software*, specially designed or adapted *course materials* and *teachers* that are familiar with the specifics of this type of education. A very important factor for the use of computer technologies is the *accessibility* of the computer system. Accessibility means the degree to which a product, a device, a service or an environment are usable by as many people as possible. In this paper we present software and hardware solutions that improve the accessibility of the computer system for students with SEN.

Computer systems in educational environments are usually standard configurations. They consist of the computer itself and a set of peripherals such as monitors, headphones, microphone, video camera, printer and scanner. These are not always enough to meet the needs of students with SEN. Depending of the type of a disability it may be necessary to use different types of assistive technologies or standard technologies adapted to the personal needs of students with SEN. The selection of the technology is highly individual and depends mainly of the type of disability.

2 Students with SEN

Depending on the type of disability, users with SEN could be generally divided in several major groups: *visually impaired*, *hearing impaired*, and users with different types of *physical and learning disabilities.* There are several major problems in these groups of learners and it is important to overcome these problems.

For students with *visual impairments* the main problems are associated with perception and processing of visual information. They may be with low vision or blind. This supposes the use of different computer technologies to access and create educational materials.

The main problem for students with *hearing impairments* is the perception of audio information. They need alternative performance of audio in various educational materials. This can be done through subtitles and a proper visual indication of sounds.

The main problems of students with different types of *physical disabilities* are associated with physical limitations of handling control and access to computer devices. These problems are mainly related to work with the keyboard and the mouse. This type of disability could cause the inability to hand write and could lead to speech problems. In these cases, computer technologies are an alternative for developing writing skills as well as communication tool for non-verbal students.

Students with *learning disabilities* may not have visible problems, but they have difficulties to acquire educational material. This group includes mainly students with *dyslexia.* The dyslexia is a reading disorder in the presence of the expected for age level of intelligence and education [1]. This problem affects many aspects of learning, it is difficult to diagnose it, and it may lead to delay in the learning process. Dyslexic learners perceive more easily information that is heard rather than read. Therefore it is useful for them to use text-to-speech technology.

3 Assistive Technologies

Assistive technologies is a broad term referring to "any item, piece of equipment, or product system whether acquired commercially off the shelf, modified or customized, that is used to increase support, maintain, or improve functional capabilities of individuals with disabilities"[1]. More generally, they are defined as devices or strategies used to cover the gap between the human capabilities and the requirements of the environment [2].

In terms of computer technology, this means products specifically designed to provide access to the computer system (including the ability to work with it) to people with physical or mental difficulties, disorders or disabilities [2].

[1] The US technology-related assistance for individuals with disabilities act of 1988, Section 3.1. Public Law 100–407, August 9, 1988 (renewed in 1998 in the Clinton Assistive Technology Act).

Assistive technologies can be divided into two categories – *adaptive* and *alternative*. Adaptive technologies modify or extend the functions of standard technologies, make them available for people with special needs. Alternative technologies are designed to replace conventional technologies.

Selection and use of various types of assistive technologies depend of many factors. The most important are the computer hardware, the operating system and the type of the user's disability.

The language is very important factor for usability of some technologies. Because of the language barrier, some of the existing assistive technologies are not usable in Bulgaria. For example, alternative keyboards with a limited number of keys are unusable in countries using the Cyrillic alphabet, as it has more letters than the Latin alphabet. In addition, some types of assistive software are adapted to work mainly with the English language.

In this paper we present only assistive technologies that are usable in Bulgaria, especially for users with impaired vision, dyslexia and problems with development of hand writing and speech.

4 Accessibility Tools

Built in modern operating systems (OS) the accessibility options are an important aspect in use of computer technologies in special needs education. This section describes the accessibility options of Windows. This OS supports a set of different settings that can be customized according to the different types of disabilities.

It is possible to use monitor for low vision students by applying settings to improve visibility of objects on the screen. By using these settings, we can change the size of bars, icons and cursors; the colors of windows, texts and backgrounds; the speed of cursor's motion or the blinking of text marker [3].

Learners with low vision can use screen magnifiers. These are programs that work as lens hovering over the page, increasing the size of objects on the computer screen by a predefined scale factor. Such a program is built in Windows [4].

The presentation of visual information to blind students can use different alternative ways:

- reading the screen content with text-to-speech software;
- using alternative devices, such as Braille displays and printers;
- duplicating text information with audio files (this requires development of additional alternative resources).

Speech synthesizer (text-to-speech) and screen reader are special software programs that can read the screen content. A *text-to-speech* program reads the screen text via a computerized voice. Unlike the screen reader, it can read out only the textual information (from a text document or a web page), but not any system information like file structure or changing windows. A *screen reader* reads everything appearing on the screen including icons, menus, text, punctuation and controls. Thus, screen readers transform the graphic user interface into audio (sound) interface [4].

SpeechLab 2.0 is a Bulgarian speech synthesizer. This program was created by a team from the Bulgarian Association for Computer Linguistics (BACL, `www.bacl.org`). It allows reading, creating and processing Bulgarian texts in different applications. It has options for various settings, and it can read selected text with appropriate intonation in various word processing programs or Web.

To access computer information visually impaired students can use Braille display (Fig. 1) and Braille printer. For entering information they can use Braille keyboard (Fig. 2). With the help of these devices text information is transformed into Braille and can be easily read by those students. *Braille displays* are tactile devices to display line by line and in the form of Braille the information presented on the computer screen. *Braille printers* are devices for printing a text document in Braille on paper [5].

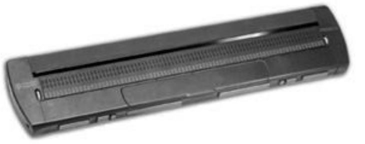


Fig. 1. Braille display

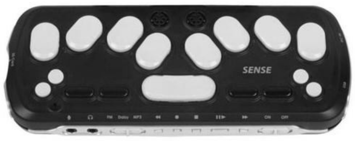


Fig. 2. Perkins style Braille keyboard

Another type of technology that can be used by students with visual and learning disabilities are the Optical Character Recognition programs (OCR). With their help, scanned text can be converted into electronic text. This process has two stages – scanning the page and recognizing the scanned image as text. Recognized text could be saved and processed by word-processing programs or could be read by using a braille display or a screen reader [4].

Visually impaired students can use as input devices like genuine Braille keyboards or standard keyboards adapted by stickers. Blind users can use Braille stickers (Fig. 3) and users with low vision – large print stickers (Fig. 4) in contrasting colors.

The mastering of the ten-fingers writing system on BDS keyboards[2] could provide a great benefit to learners with SEN, particularly those with vision impairment. This layout is optimized for writing Bulgarian language.

When the ten-fingers writing system is used, the hands are placed at specific locations over the keyboard so that each finger is “responsible” for some base key and the keys nearby. Searching positions of letters on the keyboard in that technique is not necessary – they are remembered. This allows text input without looking at the keyboard, and leads to a higher writing speed. It is especially

[2] BDS keyboard layout – Bulgarian National Standard for the arrangement of the Cyrillic alphabet letters on the keyboard.

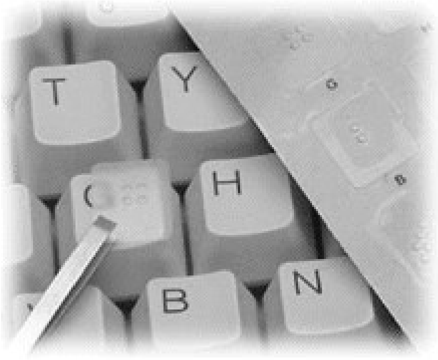


Fig. 3. Braille labels

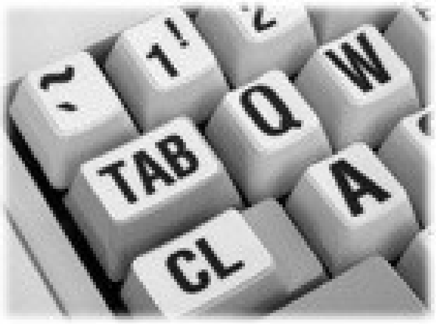


Fig. 4. Large print labels

useful for users with impaired vision. Combined with Braille stickers and speech synthesizer this technique is a great alternative for blind students.

Figure 5 shows the scheme of various technologies useful for students with different degrees of visual impairments.

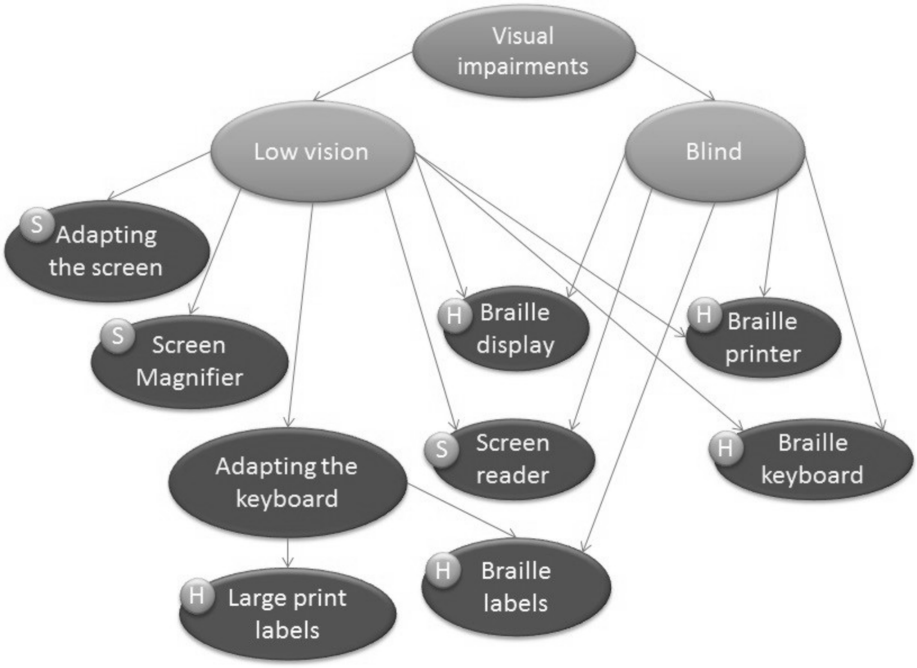


Fig. 5. Technologies for visual impaired students (H – hardware, S – software)

The accessibility options of operating systems can support physically disabled students to work with keyboard and mouse [3]. Different settings include:

- change left and right mouse buttons functions;
- change double click speed;
- select text without continuously pressed mouse button;
- change cursor shape, color, size or speed;
- automatically “sticking” mouse pointer over an object (button, icon) on the screen;
- control the mouse through the keyboard;
- activate keyboard filters for tremor, paresis and fine motor skills problems with arms, such as:
  - repeatedly pressing the same key on the keyboard (leading to multiple prints of the selected symbol);
  - holding pressed keys for a long time (also leading to multiple prints of the selected symbol);
  - pressing a few keys at the same time (leading to prints of several characters).

In these cases a *key guard* can be used (Fig. 6) to facilitate the selection of a particular key.

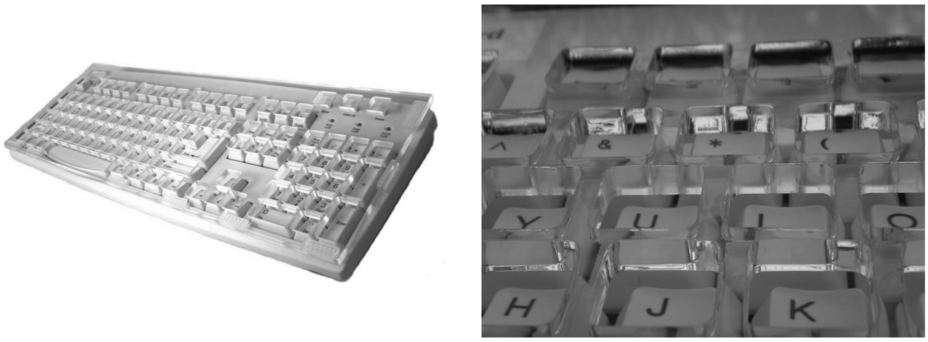


Fig. 6. Key guards for keyboards

The OS settings can facilitate the computer work of students with physical disabilities. However, this is often insufficient and requires the use of additional alternative devices and software to overcome the problems that students experience while working with a computer.

The use of a virtual keyboard can solve problems with the standard keyboard. It is a good alternative, since it can operate in different modes, and it can be controlled via the mouse or via just a single dedicated key of the keyboard.

Alternative electronic pointing devices such as *trackballs*, *joy sticks*, *switches* and others could also be used. Switches require special scanning software, which highlights objects on the screen that can be selected.

5 Making Education Accessible

The application of described technologies in education depends on the individual students' needs and the type of their disabilities. These technologies can be combined or used individually to achieve the educational purposes. The use of the computer system by students with SEN enables them to access a variety of educational resources that remain inaccessible in their original form. Access to the standard printed materials, such as textbooks is impossible or very limited for learners with visual impairments. By transforming printed text into an electronic text by using a combination of scanning and OCR technologies, these resources are now available through the computer system. Then they can be visualized with the appropriate program, read by a screen reader or Braille display, and printed on a Braille printer. This gives visually impaired students access to a variety of printed educational materials.

These technologies make the printed learning materials accessible and understandable for students with learning disabilities – particularly dyslexia. Since the main problem with such students is related to the reading, rather than the perception and understanding of information, the use of a screen reader to read the electronic text leads to overcoming of the barriers posed by reading problems. This demonstrates how a single type of technology can be used by users with different special needs.

The use of screen readers to access the information depends on the availability of the learning materials. If a text file is protected from automatic reading, it cannot be read by screen reader software. Acquiring the idea of images in the text is possible for blind users only if the images have meaningful alternative text descriptions. Therefore, very important to the learning process are not only the accessible technologies, but also the accessible educational materials.

The accessibility of learning materials is essential for students with hearing impairments. When materials are in audio format, they remain inaccessible to them. Therefore it is important for this type of materials to have alternative text versions. For example, educational video materials should be accompanied with subtitles that duplicate the speech.

Depending on the degree of which vision is affected, visually impaired learners can use Braille or modified standard keyboard for entering text. When using a standard keyboard it would be advantage to use ten-fingers writing system and BDS keyboard layout. In combination with a screen reader technologies, it allows to read each character entered by the user and helps to correct any errors in the process of writing. This set of technologies helps visually impaired learners to perform assignments that require a written answer like essays, reviews and others.

The computer keyboard can be used as an alternative tool by students with disabled fine motor skills, who cannot write by hand, but need to build the ability to write. In these cases students could use label modified keyboard with key guards and different keyboard filters.

These technologies are an alternative approach for providing new communication possibilities and skills for students, who use writing because of their inability to express themselves orally. Using a combination of text input through the keyboard and reading it through the screen reader software creates the ability to communicate with speech. It gives the students the opportunity to perform verbal tasks and to actively participate and interact in the learning process.

6 Conclusion

The technologies presented in this paper are a small part of the existing assistive technologies that could be used in special needs education. Taking into account the different problems such as language barriers, high cost of alternative devices and software we presented some of the most necessary technologies usable in Bulgaria. Furthermore, the classification of disabilities is relative. Quite often the primary type of disability causes collateral secondary disabilities of different type. This may require a combination of several technologies in order to fully address all students' needs.

Acknowledgements

This work was supported by the European Social Fund through the Human Resource Development Operational Programme under contract BG051PO001-3.3.06-0052 (2012/2014).

References

1. Kabakova, V., Boyanova V.: Dyslexia. Current hypotheses and concepts. Contemporary trends in development of speech therapy theory and practice of the 21 century, Varna (1999)
2. "ATTRAIN (Assistive Technology Training) – Assistive Technology Consultant/Advisor Training Development and Delivery". Proceeding Socrates Grundtvig Project ATTRAIN, VIENALA, Košice, Slovakia. (2004) ISBN 80-8073-230-2
3. Accessibility. Curriculum Resources for Special Education. Personalizing technology to support students with learning style differences and disabilities. Microsoft Corporation. http://www.microsoft.com/education/en-us/teachers/guides/Pages/Accessibility.aspx (23 June 2013)
4. Hasselbring, T., Williams Glaser, C.: Use of Computer Technology to Help Students with Special Needs. Children and Computer Technology, **10**(2) (Fall/Winter 2000) 102–122
5. http://www.princeton.edu/futureofchildren/publications/docs/10_02_04.pdf (23 June 2013)
6. Accessibility. A Guide for Educators. Empower students with accessible technology that enables personalized learning. Edition 3.1. Huyler, LaD., Moulton, G. (eds.). Microsoft Corporation, Redmond, Washington 98052-6399 (2011). http://www.microsoft.com/education/en-us/teachers/guides/Pages/Accessibility.aspx (23 June 2013)

Mobile Technologies in Learning

Lyubomir Serafimov

Faculty of Mathematics and Informatics, Sofia University "St. Kliment Ohridski",
5, James Bourchier Blvd., 1164 Sofia, Bulgaria
luboap10@gmail.com

Abstract. The lowering prices and the growing functionalities of the mobile technologies combined with the increasing need of interactions not only between people, but also between people and various kinds of information have led to significant broadening of the mobile technologies usage. Once used mainly for voice and messaging services, they are now tools for socialization, entertainment, shopping, work, and learning. The application of mobile technologies in the learning process was regarded a novelty 10 years ago, but now it is more and more widely considered common, and, in numerous cases, very important. The purpose of this paper is to present an overview of some key aspects of the mobile technologies applications in the learning process – brief history and definitions, current state, and tendencies.

Keywords: mobile learning, mLearning, mobiles, mobile devices

1 Introduction

The benefits of applying mobile technologies in the field of education and, more generally, in any learning process, was recognized about three decades before any mobile devices appropriate for learning purposes became available [1]. Such devices appeared with the announcement of PalmPilot (1996) and Apple's Newton, which version eMate 300 (1997) was targeted for using in the schools [2]. Along with their appearance the first articles about using mobile technologies for educational and learning purposes were published [3]. Soon after – in 2000 – the first term for denoting this use appeared – mobile learning or mLearning – and the first attempt to define it formally was made [4].

In the beginning of the 21st century were announced the first initiatives encouraging the use of mobile technologies in the learning process [5] and the first international large-scale project focused on this use – MOBIlearn in 2001.

In the following years and nowadays application of the mobile technologies for educational and learning purposes flourished, stimulated by their rapid development. Various definitions of the term mLearning appeared and it became widely accepted. The number of mLearning articles, books and initiatives quickly rose.

This paper makes an attempt to outline the fundamental issues, concerning the field of mLearning – brief history and definitions, current state, challenges and tendencies.

2 Brief History and Definitions

2.1 First Visions for Using Mobile Technologies for Learning

In his presentation "A Short History of Mobile Learning" Mike Sharples presents two of the first visions for using mobile devices for learning purposes [1]. The earlier one is of Aldiss – "miniputers" [6] and the second one – of Kay – "Dyna-Book" [7].

In the vision of Aldiss no particular learning applications are mentioned, except for the easy access to large amount of information. The Kay's vision is considered the first extensive and scientifically well-founded project for a mLearning platform [8]. Kay not only describes the technical specifications of the "Dyna-Book", but also suggests pedagogical approaches to its learning application [7].

2.2 First "Learning-Friendly" Mobile Devices and Definitions of mLearning

The first mobile devices suitable for learning purposes, were the series PalmPilot of Palm in 1996 and the Apple's Newton, which version eMate 300 (1997), was designed specifically for application in the schools [2]. The first articles about using mobile technologies for learning were also targeting schools and suggested application for learning was mainly collecting and analysing of data outside the school [3].

The first scientifically well-grounded term and definition for the utilization of mobile technologies for learning purposes appeared in 2000: "mLearning is the intersection of mobile computing and elearning: accessible resources wherever you are, strong search capabilities, rich interaction, powerful support for effective learning, and performance-based assessment. elearning independent of location in time or space." [4]. In the same article Quinn suggested a short, "working definition" of mLearning – "using Palm as a learning device".

2.3 mLearning in the First Decade of the XXI Century

In this period the number of mobile subscriptions worldwide increased over five times – from under 1 billion in 2001 [9] to over 5 billion in 2010 [10], and the monthly mobile data traffic in 2010 – 237 petabytes – was three times greater than the total internet traffic in 2000 – 75 petabytes [11]. This penetration of the mobile technologies, due to their quickly increasing capabilities and decreasing prices, along with the growing need of socialization and place and time independent learning, are the main factors that lead to the advance in the mLearning field in this period [2, 12].

The first international large-scale project, dedicated to explore "new ways to use mobile environments to meet the needs of learners, working by themselves and with others" [13] – MOBIlearn – commenced in 2001-2002, was considered "highly innovative and unusual", and the term mLearning was hardly known at this time [14]. Four years later, mLearning was still characterized as a "new concept" [15], "difficult to define, conceptualize and discuss.", but it was also considered "growing in visibility and significance" [16]. Numerous facts attest to the grown interest towards mLearning. Until 2005 were launched 10 events, dedicated to mLearning [16]. Also, about 5 institutions, 15 journals, and 15 projects were created and about 15 leading specialists in educational technologies were working in the field of mLearning [17].

A short chronological review of the main mLearning events in the period 2001–2009 is made by Gary Woodill [2]. The most important of them are: MOBIlearn project (2001-2002); bestowing over 100 awards "Palm Education Pioneer" on USA teachers (2001); the Handheld Learning Resource project in UK for encouraging the use of mLearning in extracurricular school activities (2002); establishment of the first mLearning conference – mLearn – in 2002 (at the University of Birmingham); distributing PDA devices among teachers in 27 UK schools (2002); selling of the first software product – Learning Mobile Author by Hot Lava Software – dedicated to creating mLearning course content (2005); initiation of the first international mLearning association – The International Association for Mobile Learning – at the sixth mLearn conference (2007); increasing of the number of software vendors selling software for development of mLearning course content (2005–2010).

In the beginning of the decade the mLearning was still focused on the importance of the used device, as can be noticed in the first mLearning definition for this period – that of Nantel in 2001. He defined mLearning as "a new way to learn using small, portable computers such as personal digital assistants (PDAs), handheld computers, two-way messaging pagers, Internet-enabled cell phones, as well as hybrid devices that combine two or more of these devices into one." [18]. This definition closely resembles the "working definition" of Quinn and limits the scope of mLearning to learning with small-sized devices. It was soon realized that such techno-centric definitions of mLearning are unstable, because of the rapid pace of technological development, so more user-centric approach should be applied [16]. In 2005 appeared the first definition of mLearning in which the mobility of the user is considered equal to the mobility of the technologies: "Any sort of learning that happens when the learner is not at a fixed, predetermined location, or learning that happens when the learner takes advantage of the learning opportunities offered by mobile technologies." [19]

The advance of mLearning in this decade lead to the need of software tools for its development. As mentioned above, the first software product dedicated to creating mLearning course content was developed in 2005 by Hot Lava Software. In five years only the number of such tools grew to over 30 and more than 20 articles dedicated to them were published [20]. Despite this number of software tools, their features were limited to creating mLearning content and testing the learners. The development of mLearning software tools, providing more extensive set of features, began in 2010, when the software vendors of existing eLearning software systems, started to include functionalities in their products not only for creating, but also for other mLearning activities, e.g. design, testing, delivery, assessment, management, security.

3 Current State, Challenges and Tendencies

The increasing of the capabilities and the decreasing of the prices of the mobile devices and the mobile internet access, which speed, on its side, is continuously growing, continued in the current decade [11, 21–24]. In the end of 2011 the number of worldwide mobile subscriptions reached 5.9 billion [23], and, according to some sources [24], surpassed 6 billion. In the last quarter of 2012 the smartphone sales reached 44–45% of all mobile phones sales and are expected to reach over 50% in 2013 [26]. The monthly mobile traffic grew to 885 petabytes in 2012 [22] – over three times more than this traffic in 2010 – 237 petabytes [11].

The fast technological progress and the lowering prices support the penetration of mLearning. Its application became increasingly popular not only in the educational, but also in the corporate [2, 27] and military sector [27]. The sales of mLearning software products also increased [2, 27], and appeared numerous open courses and webinars from leading mLearning specialists.

The most common applications of mLearning nowadays are [2, 12, 27–30]: audio and video podcasts, using mobile applications for the studied subjects, receiving institutional and courses news, learning assessment, collecting and analysing data in or out the learning premises, augmented reality.

The current key challenges to mLearning are [2, 12, 27, 29, 30]: lack of policy support, difficulties in changing the established institutional practices, lack of systematic pedagogical approach, negative social attitudes to the mobile devices, because of their use for entertainment and cheating purposes, copyright concerns by the publishers about the digitalising of their textbooks and the great variety of mobile devices.

Some of the mLearning tendencies determining its near future are [2, 12, 27–30]: increasing the interactivity in the learning community, learning personalization, achieved through the principle "bring your own device" (BYOD) and by developing new pedagogical approaches and improving the adaptivity of the content to the users; improving and developing new mLearning software products (e.g. mobile simulations and serious games); supporting the learners not only in the learning process, but also when applying their knowledge and skills; improving the augmented reality experience (e.g. Google Glass).

4 Mobile Learning in Europe

Europe has been a key force in the development of the mLearning since its first years. The first international large scale mLearning project – MOBIlearn, mentioned in Sect. 2.3, was a European initiative. Numerous mLearning initiatives have been implemented in the following years. One of the most comprehensive studies of these initiatives in EU was carried out by UNESCO from August-September 2011 to January 2012 and is presented in the papers [29, 30]. Only the large-scale projects are summarized here, because of their impact on the mLearning development in general.

The main funding source for European educational and scientific initiatives, including in mLearning, is EU, chiefly through its Framework Programmes for

Research and Technological Development (FPs) and other programmes, e.g. Leonardo da Vinci Programme, Lifelong Learning Programme, etc. Except for the EU-funded projects, UNESCO researchers also identified several nationally-funded and locally- or privately-funded European projects, but only two of them are large-scale.

Four large scale EU-funded mLearing projects are presented in [29]: HandLER, M-Learning, MOBIlearn, eMapps (here they are chronologically ordered).

The Handheld Learning Resources Project (HandLER) project ran from 1998 to 2002 in UK and targeted the development of mobile technologies and methodologies that could support lifelong learning. Although the developed technologies were significantly limited for use by a mobile learner – devices were heavy, had short battery life, etc., the project can be determined as successful, because it helped to form the concept of mobile learning outside the school settings and defined some basic requirements for mobile learning devices.

The purpose of the M-Learning programme (2001–2004) was launched to support 16–24 years old people who had not succeeded in the educational system by involving them in informal learning activities. The focus of the project was not on the delivery of learning content on mobile devices, but on encouraging creativeness and collaboration in the learning activities. The project was successful because of the reached insight that mLearning is most effective when combined with other types of learning, e.g. eLearning, and not when used unaccompanied.

MOBIlearn (2002–2005) was a large-scaled international project, involving not only nine European countries, but also USA and Australia. The initiative included three pilot projects for supporting the learner in different settings – in higher education, in museums and galleries, and in acquiring first aid skills. The main success of MOBIlearn was not the technological one – that some of the created devices and software were commercialized, but the conceptual, because it shifted the focus from the mobility of the device to the mobility of the learner, which is long-term project impact.

The eMapps (Motivating Active Participation of Primary Schoolchildren) programme (2005–2008) was aimed to 9–12 years old students in 10 countries, and its target was to form groups of creative, technologically-experienced, and cosmopolitan children who would produce digital content concerning their local culture and connect with children in other countries, as well as to develop software for mobile devices, primarily games, supporting the learning. The project was successful both technologically and pedagogically, because the created software was effectively coupled with pedagogical approaches, which actively engaged the children.

The two large-scale nationally-funded projects presented in [29, 30] is the UK programme Mobile Learning Network (MoLeNET) and the French UnivMobile.

MoLeNET ran from 2007 to 2010 and involved not only learners, but also teaching staff. Its aim was to provide various types of support to a variety of activities, comprising the use of mobile technologies in education. Benefits of the projects were not only the improvements in learner attendance, achievements, drop-out rates, and staff motivation and collaboration, but also in confirming the mLearning insights, reached in the projects M-Learning and MOBIlearn – that the mLearning is most effective when combined with other types of learning and that focus should be on the learner.

UnivMobile was launched in 2009 and is still active. Its aim is to support communication between staff and students in French universities via mobile applications. The project helps students in navigating university life – contacting professors, checking course schedules, exam dates and test results, receiving university news and lecture podcasts accessing interactive maps of university campuses. The staff is supported by easing the communication with students via mobile applications for sending class schedules, exam dates, test results, uploading podcasts.

UNESCO researchers conclude their study with the optimistic view that the number of mLearning projects in Europe will increase, because powerful mobile devices are quickly disseminating not only among the students, but also among the staff, which will encourage more countries to participate in mLearning initiatives.

5 Conclusion

Mobile technologies are stimulating changes in the way of learning because they allow it to be truly time and place independent, more personalized, interactive and just-in-place/time, and, due to their decreasing prices, more accessible both to the institutions/companies and the learners. That is why the mobile technologies had, have and will have an important role in encouraging the dissemination of knowledge.

Acknowledgements

This work was supported by the European Social Fund through the Human Resource Development Operational Programme under contract BG051PO001-3.3.06-0052 (2012/2014).

References

1. Sharples, M.: A Short History of Mobile Learning (2007)
2. Woodill, G.: The Mobile Learning Edge: Tools and Technologies for Developing Your Teams (1st ed.). The McGraw-Hill Companies, Inc. (2011)
3. Curriculum Administrator: Distributed learning environment opens door to lifelong learning. Curriculum Administrator **31**(6) (1997)
4. Quinn, C.: mLearning: Mobile, Wireless, In-Your-Pocket Learning. LineZine, Fall (2000) http://www.linezine.com/2.1/features/cqmmwiyp.htm
5. PEP (Palm Education Pioneer Grants): SRI International (2002) http://palmgrants.sri.com/ideabank.html (last accessed on 1 of August, 2013)
6. Aldiss, B.: The thing under the glacier. In: Daily Express Science Annual No. 2 (1963)

7. Kay, A.: A Personal Computer for Children of All Ages. In: ACM '72 (Proceedings of the ACM annual conference) Vol. 1 Article No 1 (1972)
8. Sharples, M.: Disruptive Devices: Mobile Technology for Conversational Learning. International Journal of Continuing Engineering Education and Lifelong Learning **12** (2003) 504–520
9. Plunkett, J.: Plunkett's Telecommunications Industry Almanac 2003–2004. Plunkett Research Ltd. (2002) ISBN-13: 978-1891775222
10. Plunkett, J.: Plunkett's Telecommunications Industry Almanac 2011 (1st ed.). Plunkett Research Ltd. (2010) ISBN-13: 978-1593921781
11. Cisco: Global Mobile Data Traffic Forecast Update, 2010–2015 (2011)
12. Quinn, C.: Designing mLearning: tapping into the mobile revolution for organizational performance (1st ed.). San Francisco, CA: Pfeiffer, An Imprint of Wiley (2011)
13. MOBIlearn: MOBIlearn Project – Vision (2002) `http://www.mobilearn.org/vision/vision.htm` (last accessed on 1 of August, 2013)
14. Attewell, J.: Mobile technologies and learning: A technology update and m-learning project summary. Learning and Skills Development Agency (2005)
15. Kukulska-Hulme, A., Traxler, J.: Mobile Learning: A Handbook for Educators and Trainers (Open and Flexible Learning Series) (1st ed.). London, Routledge (2005)
16. Traxler, J.: Defining Mobile Learning. IADIS International Conference (2005) 261–266 ISBN-10: 972-8939-02-7
17. Cobcroft, R.: Literature Review into Mobile Learning in the University Context. Queensland University of Technology, Creative Industries Faculty (2005) `eprints.qut.edu.au/4805/01/4805.pdf`
18. Nantel, R.: How to Determine Your Readiness for Mobile e-Learning. Brandon-Hall (2001)
19. O'Malley, C., et. al.: WP4 – Pedagogical Methodologies and Paradigms (2005) `www.mobilearn.org/download/results/public_deliverables/MOBIlearn_D4.1_Final.pdf`
20. Mugwanya, R., Marsden, G.: Mobile Learning Content Authoring Tools (MLCATs): A Systematic Review. In: Villafiorita A., Saint-Paul, R., Zorer A. (eds.). Proceedings of the 1st conference on E-Infrastructures and E-Services on Developing Countries, Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, Maputo, Published by Springer Berlin Heidelberg (2010) 20–31
21. Cisco: Global Mobile Data Traffic Forecast Update, 2011–2016 (2012)
22. Cisco: Global Mobile Data Traffic Forecast Update, 2012–2017 (2013)
23. ITU: The World in 2011 – ICT Facts and Figures (2011)
24. Chetan Sharma Consulting: State of the Global Mobile Industry (2011)
25. Gillet, J.: Global mobile connections to surpass 6 billion by year-end (2011) `https://gsmaintelligence.com/analysis/2011/09/global-mobile-connections-to-surpass-6-billion-by-year-end/299/`
26. Gartner: Market Share Analysis: Mobile Phones, Worldwide, 4Q12 and 2012 (2013) `http://www.gartner.com/id=2334916`
27. Atkins, S.: The US Market for Mobile Learning Products and Services: 2010–2015 (2011)
28. Quinn, C.: The Mobile Academy: mLearning for Higher Education. Jossey-Bass, San Francisco. (2011) ISBN-13: 978-1-118-07265-3
29. UNESCO: Turning on Mobile Learning in Europe (2012)
30. UNESCO: Mobile Learning for Teachers in Europe (2012)

Overview of Emerging Digital-Real World Technologies in the Context of Learning Theories

Albena Antonova

Faculty of Economics and Business Administration,
Sofia University “St. Kliment Ohridski”,
125, Tsarigradsko chaussée Blvd., Bl. 3, 1113 Sofia, Bulgaria
a_antonova@fmi.uni-sofia.bg

Abstract. Technologies continue to evolve and to transform educational and life-long learning landscape. Soon lecturers and students will be confronted to totally new instruments and applications, leaving the computer screens and merging digital world with the real world. For example augmented reality, 3D printing, Internet of Things and ubiquitous computing are instruments for new complex digital-real world interactions, extending the world of ideas, text and digital information and displaying it in the world of real things, experiences and contexts of interactions. One of the main questions today is how new technologies will contribute to the process of knowledge acquisition and learning? How we will adapt learning solutions from mixing digital content to physical and real world objects and experiences? The present research will investigate the basic learning theories and analyze how emerging technologies, making the bridge between digital and real world can add value and contribute to the knowledge construction and learning. Through reflections and different examples of educational applications there will be summarized some of the main challenges and opportunities for adopting emerging technologies in learning.

Keywords: emerging technologies, technology enhanced learning, mobile application, learning theories

1 Introduction

Information technologies have largely influenced and transformed the way people learn, socially interact and construct new knowledge. Moreover, Internet profoundly changed the learning patterns and educational practices worldwide as it became the primary source of reference, content-rich, independent and easily-accessible universal platform for services. Internet extended communication opportunities and social interactions and reversed the roles of educators and learners. However, observing new technology trends we can expect that we are still in the beginning of the educational landscape transformation. Nowadays the main e-learning applications in Internet are mainly based on e-learning platforms and content-delivering services, complex simulations, social Internet and web 2.0 communities, virtual worlds, social serious games, experimentations and others. However, many researchers and practitioners anticipate profound impact of new coming ubiquitous technologies that will merge digital and real world as for example augmented reality (AR), 3D printing and Internet of Things (IoT). They expect that all these technologies will contribute to development of new content- and context-rich connected environments and objects, adding many of the Internet functions outside the computers to the "real world" objects, people and landscapes. With fast adoption of smart phones and tablets, interactive technologies, wireless Internet and portable devices (as AR glasses or head-mounted devices), this vision is gradually becoming reality. How can we prepare and adapt to these new coming challenges? Many researchers and experts are using different methods to predict the future, providing different scenarios and roadmaps (TEL-MAP). Our approach in the present research is bottom-up and theory based. As new smart and ubiquitous technologies are fast reaching our pockets, how can we use them for learning purposes? As these technologies become more complex and have this capacity to transfer objects from digital to real world (as 3D printing) or to extend the real world with digital information, as augmented reality, we need appropriate research of learning theories and approaches. That is why the knowledge construction and learning will be analyzed as central concept in the classic learning theories. This way the paper will provide reflections and discussion about the new challenges of technology-enhanced learning and its impact on the future educational landscape. The first part of the paper will make a short review of most relevant learning theories, learning models and concepts of learning. Then will be provided a short overview and examples of some of the key emerging "bridge" technologies connecting digital content and Internet to the real and physical world. The second part aims to introduce and to analyze the term educational validity of new technologies. Finally there will be discussed several examples and reflections of how new applications can be applied in learning, identifying the impact of new ubiquitous technologies on the learning process.

2 Background

2.1 Learning and Learning Theories

Learning theories have been evolving since Antiquity and reflect different stages, concepts and understandings about "what is knowledge" and "how knowledge is acquired". The processes of learning and knowledge acquisition are complex and are discussed in different disciplines as philosophy, cognitive and group psychology, sociology, informatics and many others.

Naturally, people associate learning to formal education processes and practices. Interestingly Sawyer [13] highlights that contemporary mass educational institutions and schools have been designed in 18 and 19 century, long before studies and research investigations about the specific aspects of how people learn. Thus, for traditional classroom practices (instructionism), knowledge is considered as collection of facts and procedures, and the goal of educational system is to transmit these facts and procedures to the learner. Nowadays, the research on learning and knowledge acquisition is progressing and proposes many approaches about how to conceptualize and organize learning in order to make it more efficient, effective and flexible. In the context of e-learning and technology-enhanced learning are largely investigated psychology-based learning theories (Hammond [5]), including behaviorism, cognitivism, constructivism and social constructivism (Mödtritscher [9], Hung [6], McLeod [8]). Some other researchers select as well the theory of "connectivism" and even come to "eclectism" to summarize that the modern learning theories are going to convergence. Recent studies as for example De Jong [2] are exploring new approaches to learning – "Learning by Design", combing both principles of inquiry and collaborative learning, including learning about the domain knowledge, learning about the inquiry process, and learning about cultural aspects and cultural differences. Although these new-emerging approaches, the basic learning theories can best illustrate how the process of learning and knowledge acquisition can be approached and understood as complex and context related matter.

Our analysis will further identify the main characteristics of learning theories that can help us later to better understand the role of new emerging technologies to enhance learning and knowledge acquisition.

- *Behaviorism* views learning as a function of external stimulation of the learner, who responds on external positive or negative stimulus. Some of the main contributors and supporters of this theory are J. Watson, I. Pavlov, B.F. Skinner, E.L. Thorndike, Bandura and others. The learner is essentially passive, the learner's mind is "black box" and the preliminary knowledge is not important ("*tabula rasa*"). This way the process of learning is based entirely on external factors. Behaviorism has been proved useful for development of some types of skills – especially those that can be learned substantially through reinforcement and practice.
- *Cognitivism* explores the cognitive capacity of the human mind and its ability to acquire knowledge. It investigates the functions of the memory, thinking and reflecting and it is connected to previous experience and existing knowledge. Moreover, the learner process information according to his unique learning and cognitive style. Some of the main contributors include Gagne, Briggs, Wager, Bruner, Schank, Merrill, Reigeluth, Scandura and others. Knowledge is interpreted as a schema or symbolic mental constructions and therefore learning is defined as change in a learner's schemata.
- *Constructivism* is based on the assumption that learners construct their own reality, and the individual's knowledge comprises a complex model of prior knowledge and experiences, mental structures, reflection mechanisms to interpret objects and events. Learning is an active, contextualized process of constructing knowledge rather than acquiring it. People actively construct or create their own subjective representations of objective reality. New information is linked to prior knowledge, and thus mental representations are subjective. Originators and important contributors are Piaget, Dewey, Vico, Rorty and Bruner. Each person has a different interpretation and construc-

tion of knowledge process. The learner is not a blank slate (tabula rasa), but brings past experiences and cultural factors to a situation.
- *The social constructivism* is based on the Vygotsky work, highlighting the role of the social environment and culture for reconstruction of knowledge. This way knowledge is constructed based on personal experiences and hypotheses that are supported in the environment. Learners continuously test these hypotheses through social negotiation.
- *Connectivism* aims to explore the principles of life-long learning, stating that the process of learning is not strictly fixed and occurs within nebulous environments. Learning can reside outside of learners (within an organization or a database) and is focused on connecting specialized information sets, and the connections that enable them to learn more are more important than the current state of knowing. Thus knowledge that resides in a database needs to be connected with the right people in the right context in order to be classified as learning.

Learning theories have been considered not as competing, but as complementing approaches, combining in order to reflect specific learning situations and context. Moreover, all of the theories have advantages and disadvantages, and reflects different learning specifics and contexts. Illeris [4] proposes a comprehensive framework that combines the variety of learning theories appropriate for adult learning. The proposed framework is based on three main learning dimensions: cognition, emotion and environment, and is situated in a social context. Therefore, he makes an attempt to differentiate the adult learner as one who controls the learning situation.

Besides on learning theories, researchers have been working on many different visions and concepts of learning and knowledge acquisition as: active learning, learning with understanding, pre-existing knowledge, meta-cognition (and self-assessment) (Bransford et al. [1]), experiential learning (Kolb [7]), the ownership of the learning, formal and informal learning and others. There were identified some important findings, reflecting organization of learning and knowledge acquisition (Bransford et al [1]):

- preconceptions and prior knowledge play an important role in educational process,
- building expertise requires an accumulation of factual knowledge, its good understanding and its organization into usable knowledge;
- the meta-cognitive approach of instruction allows students to take control of their knowledge, to set learning objectives and to follow specific learning paths.

The present research will further analyze and demonstrate how new bridge technologies between digital and real-world can contribute and add value to the learning theories and the learning process.

2.2 Learning Tools Validation

Learning tools need to be properly validated, in order to be justified as credible and effective mechanisms for transferring knowledge, applied in appropriate learning context. Educational validity designates how learning tools contribute to the acquisition of knowledge. Discussing the educational role of interactive business simulations, Staiton et al. [14] identify a validation framework for learning technologies. In this framework there are analyzed two different settings:

- Internal validity – learners should be able to learn and "acquire knowledge" using the learning tools;
- External validity – learners should be able to relate their educational experience (acquired with learning tool) to other real-world contexts and situations.

Applying this framework on the new emerging technologies can help us not only to understand the mix between digital and real-world applications, but what is their educational value for internal and external validity.

2.3 Defining e-learning, m-learning and Ubiquitous Learning

The recent popularity of e-learning is a consequence of wide adoption of information technologies and Internet. E-learning has been defined as complementary channel of communication allowing computers and computer networks to connect learners with learning media, learners with other people, learners with data and learners with processing power (Morrison [10]). E-learning technologies can perform customized, cheaper, flexible and learner-oriented training, reflecting personal attitudes and allowing new type of learning process, fostering significant improvements in accessibility and opportunity to learn (Teo and Gay [15]). The mobile learning or m-learning is a sub-class of e-learning technologies adapted to mobile devices. Peng et al. [12] investigated m-learning in literature, summarizing its most important characteristics: mobility, ubiquity, "anytime-anywhere" learning concept, convenience and infrastructure. The proposed definition of m-learning is: '*in order to benefit from convenience, expediency, and immediacy, mobile learners use ubiquitous computing technologies to learn the right thing at the right time at the right place*' (see Peng et al. [12]).

Ubiquitous computing was coined as a term by Mark Weiser [16] in 1991. He stated that "the most profound technologies are those that disappear. They weave themselves into the fabric of everyday life until they are indistinguishable from it." (Mark Weiser [16]). Ubiquitous computing consists of mobile devices, wireless networks and other advanced technologies and infrastructure. Moreover, it can be expected that ubiquitous learning will be reintegrated and complement the functionality of different "smart" objects such as smart phones, smart boards, smart TV sets and smart walls. Thus some researchers as Muntean and Muntean [11] state that learning should be defined as ubiquitous learning, where technologies provide continuous and context-based educational material to learners anytime, anywhere, and from any device.

Among ubiquitous technologies that will be discussed more in details in this paper are those that makes the bridge between digital world and the physical real world. The selected technologies are ubiquitous and mobile, and recently have acquired substantial interest from learning community and practitioners. Moreover, these technologies have already viable commercial applications and there are performed many initial researches and prototypes that explore different learning applications.

- Augmented reality technologies extend and mix real and virtual objects. It allows us to acquire simultaneously digital content, Internet and computational services to real life tasks, contexts, landscapes and people.
- Internet of Things (IoT) refers to emerging trend of augmenting physical objects and devices with sensing, computing, and communication capabilities, connecting them to form a network and making use of the collective effect of networked objects (Guo et al. [3]). The IoT requires the learner to orchestrate its personal intelligent eco-system and can provide platform for knowledge acquisition and experimentation.
- 3D printing is a new emerging technology that allows users to turn any digital file into a three dimensional physical product. It opens a new floor to imagination and creativity, helping us to better understand the construction of the physical world.

3 Emerging Technologies and Learning Theories

Emerging technologies as AR, 3D printing and IoT can be used as learning tools that can support and enhance traditional learning theories (Peng et al. [12], McLeod [8], Mödtritscher [9]). AR learning applications contribute to improve acquisition of factual information and understanding of physical objects. Therefore, it can be applied in the framework of traditional learning, or to use it as explorative and experience-based learning scenarios. IoT applications can improve both adaptability of learning tools and learning environment, providing feedback, customization and networking capacity. Some of the 3D printing applications allow experience-based and explorative learning by-doing. Some of the most relevant implementations of ubiquitous learning technologies are summarized in Table 1.

3.1 Examples of Emerging Ubiquitous Learning Technologies

There already exist many applications of emerging ubiquitous technologies that can be implemented in different educational context. It should be stated however that these applications are mainly pilot experiments. For example, AR applications have been implemented in archeology, architecture, art, complex learning simulations, virtual collaboration and virtual classes, virtual textbooks and educational 3D models, tourism and space-related information, translation, entertainment, AR cinema and TV applications and many others. 3D printing

Table 1. Summary of possible ubiquitous learning applications inside key learning theories

	AR – *mix of real and virtual objects*	IoT – *networked objects*	3D Printing – *prototyping*
Behaviorism	Increase factual information in real-world environment	Provide active learning environment based on feedback and response	Prototyping and construction of physical objects
Cognitivism	Improve understanding via complex information visualization	Adapt and customize learning tools and environments to increase cognitive processes	Improve cognitive processes with reconstruction of physical objects
Constructivism	Mix of virtual representations and versions of the real world; explorative learning, scenarios-based learning;	Learning tools and environments are customized and adapted to previous experience and knowledge	Improve experience-based learning, reconstructing own models and objects
Social constructivism	Increase access to social learning and shared experiences, social information and shared experiences	Increase social interaction on learning environments and tools.	Improve sharing and collaboration on physical objects designs; social networking;
Connectivism	Improve connections between physical and virtual sources, people and sensors	Improve context-related content and services, related to sensor information	Connect physical objects building to social networks and collective experience

is still a costly infrastructure, but researchers and practitioners actively work on applications in design and industrial design applications, building, fashion, food, medical applications, bio-printing and others. The IoT applications are also evolving, aiming to integrate different systems and objects inside interconnected environments. Some pilot applications have been recently explored as Pling Plong® (pillow for children, reading books) or Nabaztag® (device reading information from Internet).

4 Discussion

While there can be identified many possible applications of emerging digital-real world "bridge" learning tools, it can be assumed that soon educators will find new ways to enhance and improve their learning practices with these technologies. Ubiquitous technologies can contribute and add substantial value to the learning process, improving previous knowledge and pre-requisite: adding context and adapting to learning objectives; improve understanding and acquisition of factual knowledge; improve active participation in learning and metacognition (Bransford et al. [1]). Emerging ubiquitous learning technologies aim to reinforce the vision of learning "anytime anywhere" and to contribute to more efficient learning process. One of the substantial differences with other Internet and ICT applications is that the ubiquitous technologies are part of the physical environment and everyday real-life practices. This gives better opportunity to connect learning to real-life problems, objectives and goals and to lead to real-life experiences, allowing learners to increase both understanding, ubiquitous access to context-related data, social interactions and physical exercises. It is important as well that new emerging technologies, merging the digital and real world as Augmented Reality, Internet of Things and 3D printing need to be designed and applied in educational context taking into account both internal validity – as mediums and learning tools for transmitting new knowledge and external validity – as vehicles, enabling learners to export this knowledge to other general real-world contexts.

5 Conclusions

While it is not obvious which technologies will persist and will become a game-changer in the TEL field, it is clear that implementation of new technologies in education is a complex process. It can be expected that ubiquitous learning together with ubiquitous learning technologies would be implemented gradually, bottom-up and from informal-to-formal learning. However, many practitioners and researchers claim that emerging technologies in the field of AR, IoT and 3D printing soon will have significant impact on society, on learning and knowledge acquisition. Thus, learning experts have to be prepared how to introduce and implement this variety of technologies for specific educational settings, especially in the BYOD perspective. Moreover, adoption of new technologies in learning has to be both economically justified and to bring substantial value to the learning process. Another factor for introducing innovative technologies are educational institutions, as it is related to additional resources, funds, time and training, that have to be part of the overall TEL strategy.

References

1. Bransford, J., Brown, A., Cocking, R. (ed.) How People Learn, Brain, Mind, Experience and School. National Academy Press, Washington D.C. (2000)
2. De Jong, T.: Learning by Design. In: Wild, Lefrere, Scott (eds.). TEL2020: Technology and Knowledge in the Future, a Roadmap (2013)
3. Guo, B., Yu, Z., Zhou, X., Zhang, D.: Opportunistic IoT: Exploring the social side of the internet of things. In: Gao, L., Shen, W., Barthès, J.-P.A., Luo, J., Yong, J., Li, W., Li, W. (eds.). Proceedings of the 2012 IEEE 16th International Conference on Computer Supported Cooperative Work in Design (CSCWD), IEEE (2012) 925–929

4. Illeris, K.: Towards a contemporary and comprehensive theory of learning. International Journal of Lifelong Education **22**:4 (2003) 396–406
5. Hammond, L., Austin, K., Orcutt, S., Rosso, J.: How People Learn: Introduction to Learning Theories. In: The Learning Classroom: Theory into Practice. A Telecourse for Teacher Education and Professional Development. Stanford University School of Education, Stanford University (2001) 22 pp. `www.stanford.edu/class/ed269/hplintrochapter.pdf`
6. Hung, D.: Theories of learning and computer mediated instructional technologies. Educational Media International, Routledge `http://www.tandf.co.uk/journals`
7. Kolb D.A.: Experiential learning, experience as the source of learning and development, Englewood Cliffs, NJ: Prentice Hall (1984)
8. McLeod, G.: Learning Theory and Instructional Design. Learning Matters **2** (2003) 35–43
9. Mödritscher, F.: e-Learning Theories in Practice: A Comparison of Three Methods. Journal of Universal Science and Technology of Learning **0**(0) (2006) 3–18 `http://www.jucs.org/justl_0_0/elearning_theories_in_practice`
10. Morrison, D.: E-learning strategies, Wiley &Sons, Chichester (2003)
11. Muntean, C.H., Muntean, G.M.: Open corpus architecture for personalised ubiquitous e-learning. Personal Ubiquitous Computing **13**(3) (2009) 197–205
12. Peng, H., Su, Y.J., Chou, C., Tsai, C.C.: Ubiquitous knowledge construction: mobile learning re-defined and a conceptual framework. Innovations in Education and Teaching International **46**(2) (2009) 171–183
13. Saywer, R.K.: The Cambridge Handbook of the Learning Sciences. Cambridge University Press (2006)
14. Staiton, A., Johnsons J.: Educational Validity of Business Gaming Simulation: A Research Methodology Framework. Simulatin & Gaming **41**(5) (2010) 705–723
15. Teo, C.B., Gay, R.K.L.: A Knowledge-Driven Model to Personalize E-Learning. ACM Journal on Educational Resources in Computing 6(1) (March 2006) Article No. 3
16. Weiser, M.: The Computer for the Twenty-First Century. In: Scientific American (1991) 94–10

Knowledge Management in NABOO

Mila Dragomirova

Faculty of Mathematics and Informatics, Sofia University “St. Kliment Ohridski”,
5, James Bourchier Blvd., 1164 Sofia, Bulgaria
dragomirova.mila@gmail.com

Abstract. Knowledge and innovation are widely acknowledged as key drivers for sustainable development and competitiveness world-wide. Small and medium enterprises (SMEs), however, have large difficulties in their access to up-to-date knowledge. Therefore, they use various knowledge services offered by professional organisations in the field or enter virtual communities where they could communicate with their peers and gain knowledge from the community. Optics is a specialised sector taking advantage of new technology developments in the field, and thus, experiencing a high pace of change. This sector in Bulgaria is dominated by SMEs, mainly small and micro companies, which gain targeted assistance by the Bulgarian National Association of Optometrists and Opticians (NABOO). The organisation, however, faces several problems for serving better its members, and needs a specialised knowledge management tools. The objective of the paper is to present a case study of knowledge management in a community of practice and to propose a solution for overcoming the existing problems which NABOO presently faces.

Keywords: knowledge management, branch organizations, communications platform

1 Introduction

Nowadays, knowledge and innovation have become key drivers for sustainable development and competitiveness world-wide. The rapid development of science and technology has faced organizations with many challenges and the need to continuously monitor their environment, in particular their clients’ needs, competitors’ behaviour and the regulatory amendments. This requires a continuous learning of the organizations and increasing their organizational knowledge [1]. Subsequently, for small and medium enterprises (SMEs) it is often impossible to follow the development trends in their field, as well as to continuously improve their organizational knowledge capabilities. Thus, their participation in professional organizations, online communities, and the so-called process of embedded innovation facilitates meeting these challenges. As noted in [2], the generation of innovation in SMEs is related to multilateral relationships – they collaborate with other SMEs, scientific communities and other organizations, from which they obtain up-to-date knowledge and expertise, share and discuss ideas and concepts.

Therefore, knowledge management (KM) scholars pay special attention to virtual communities, which generally are based on common goals, interests, beliefs, and activities, and follow common rules for interaction and sharing knowledge using common information space [3]. Unlike virtual communities, sectoral and professional organizations really work with their members, and in the process of their development accumulate a variety of knowledge that they share with their members or use to achieve joint objectives.

The goal of this paper is to present a case study of a Bulgarian professional organization and the possible way for solving the problems it faces with KM and better serving its community. The paper follows a Swiss methodology for case studies writing [6], which was transferred by the authors to Sofia University staff within recent projects. The methodology includes the following phases:

- General information, vision and goals of the organisation, and the importance of information technology (IT) for its KM strategy;
- Launching a project – reasons and problems faced;
- Description of the specific solution;
- Project implementation, including management, software development and maintenance;
- Experience obtained during the implementation;
- Success factors – linked to the specific solution and its practical implementation.

In this particular case, the project for solving the problems of the organization is in initial phase. Therefore, the paper focuses only on the initial parts of the methodology, and in the conclusion considers the expected benefits, and the possible success factors and the way ahead.

2 NABOO Key Activities and Challenges Faced

The Bulgarian National Association of Optometrists and Opticians (NABOO) was established in 1998 in response to the needs of the ophthalmic optics community, and with the aim to protect their interests [4]. Members of the organization are individuals working as optometrists or opticians, companies operating in the field of ophthalmic optics, optical shop owners, manufacturers, importers of optical or contact lenses and spectacle frames, and suppliers of equipment and processing tools glasses. In the organization can join also schools that provide education in the field of optometry and ophthalmic optics. The activities of NABOO include:

- consulting, legal and economic assistance and information services to its members;
- development of standpoints, and assistance on matters concerning the normative base, the politics, the research and education in the field;
- development of standards and norms for professional behaviour and ethics in the sector;

- organizing seminars, workshops and meetings to raise the awareness of its members on the trends and achievements in the optical industry;
- cooperation with other organizations in the country and abroad;
- support for contacts and expanding the international relations of its members;
- protection and representation of its members to relevant state, municipal, and international agencies and institutions.

As it is shown in Figure 1, NABOO maintains a major network of stakeholders in order to better perform its functions in Bulgaria.

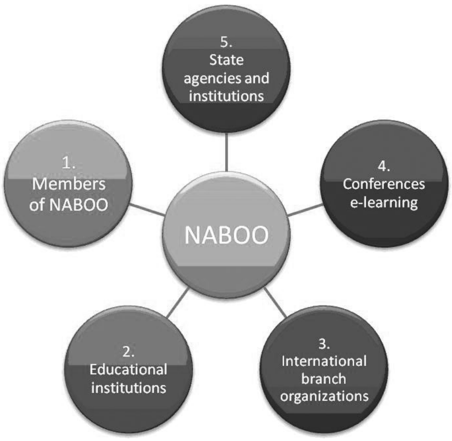

NABOO
1. Members of NABOO
2. Educational institutions
3. International branch organizations
4. Conferences e-learning
5. State agencies and institutions

Fig. 1. Main participants and contacts of NABOO

The performance of the professional tasks in this industry requires maintaining high qualified personal, able to learn-on-the-job in order to follow the trends in the sector. Thus, some of the challenges of NABOO are linked to specialized training and skills upgrade of its members in order to practice optometry and optic at highest quality. Here, some recent changes in the formal education, both at secondary and higher level, ensure the necessary skills supply. However, the dynamic development of technologies requires a constant dialogue with educational institutions for providing training according to the changing skills demands of the industry. In addition, life-long-learning and sharing of best professional practices, discussions on professional studies are facilitated by many international organizations in the field, however, are not well developed in the Bulgarian community. This could be ensured by a suitable web platform providing a common knowledge base and e-learning tools for the NABOO community.

Regulations, standards and the overall trends in the sector require monitoring of them, as well as collaboration with national authorities in the area of health and protection of competition. The exchange of information and knowledge with similar international and European organizations, as well as on a bilateral base, is also an important element of the NABOO activities. For example, the membership in the European Council in optics and optometry (ECOO) enables maintaining a wide range of contacts with leading professionals from the optical sector, and raising the public importance of the profession, using the support and assistance of experts from Europe and the world.

It is obvious that in order to perform better and to fulfil its tasks for improving the skills and competencies of the community, NABOO requires managing better the knowledge coming from its members, as well as from external sources. While the predominant part of this knowledge is explicit (documents, regulations, standards, research papers, best practices, etc.), it needs to be acquired, stored, and organized in order to be easily found and used when needed. On the other hand, a large variety of knowledge comes from communications among professionals, and sharing it with other interested stakeholders could be facilitated by a suitable communication platform accessible by the whole NABOO community.

The idea of using information and communication technologies (ICT) to improve sharing of information between the NABOO members has its history related to unsuccessful development of two websites that provided general information about the organization and its members, including contact details, news in the sector, etc. A major problem is the absence of a professional concept to develop a Website that could help to solve the information and communication needs of NABOO and its members, and to ensure its subsequent maintenance, adding new functionality and features. Using a common forum and social networks (e.g. Facebook) gives temporary results. However, the overall knowledge sharing occurs with varying degrees of activity, chaotic and disorganized and ineffective despite the enthusiasm and efforts of some members. The major problem is that there is no clear strategy how to ensure the knowledge transfer and which technologies to be used.

3 Concept of Collaborative Platform

The analysis of the present state and the needs of NABOO members clearly show that a knowledge management strategy should be developed with a focus on ensuring a common platform to facilitate organizational learning, knowledge-sharing and communications among NABOO stakeholders. The major users of such collaborative platform will be the members of NABOO, interested scientific and educational organizations, manufacturers of high-tech optical products, state and local organizations as well as citizens. The objective is the platform to become a single entry point for information and knowledge sharing in the optics sector.

A number of methods and tools could be used for the development of the platform such as: yellow pages, knowledge maps, social networks, etc. [1]:

- Knowledge maps can visualize what sources of information and knowledge (people, libraries, databases, online resources, etc.) exist and where they are available, thus helping quickly to access them.

- Yellow Pages are regularly used to locate expertise and to find experts with a specific experience, knowledge and skills. They represent a database of structured information for members and experts of the organization, and should be accessible to all stakeholders, as well as should provide opportunities to the experts for information update.
- Taxonomy as classification systems allow for clustering of knowledge and information so that they can be systematically processed and reused. The use of metadata (tags) to describe knowledge objects allows to set the knowledge context and attributes, and to provide information to be used in the operations for their search and retrieval.

Obviously one of the main features of the web platform of NABOO could be to localize expertise through the yellow pages – 'who is who':

- NABOO – aims and activities of the organization, structure, services provided documents required for membership (forms, returns);
- Members of NABOO – maintain a current list of members with contact information, specialization and more;
- educational institutions providing training in the field of optometry and optics – contacts, education and training programs, application conditions;
- partner organizations (European, global, industry and professional organizations in other countries and in Bulgaria) – a short presentation, contact information;
- governmental organizations with regard to the industry – functions and contacts.

Second significant functionality should be the organizational knowledge base, including:

- register for paid membership fees;
- normative base, regulation and performance requirements of the sector;
- educational materials;
- opportunities for collaboration with educational institutions – assistance for internships and possible scholarships from employers, joint projects;
- information on related organizations, and foreign best practices;
- access to a library of professional literature.

Informing periodically NABOO members about the news in the field is also essential, and should comprise information about events organized for NABOO members, external events (e.g. conferences, exhibitions and other events in the field of optometry and optics), decisions concerning the sector, new development trends, etc.

Very important is the web platform to enable the communications between NABOO members, and with external organizations. Providing access to a collaborative platform for various stakeholders, e.g. from industry, the scientific community, educational institutions, etc. could significantly contribute to the transfer of knowledge between them using variety of communication channels (newsletters, intranet, meetings, seminars, training, email, etc.). In practice, the goal is to create the conditions for the formation of a strong Community of practice (CoP) as an informal network of people with shared values and beliefs, and this virtual community to be supported by various web technologies for communication, knowledge sharing, and discussions (Figure 2).

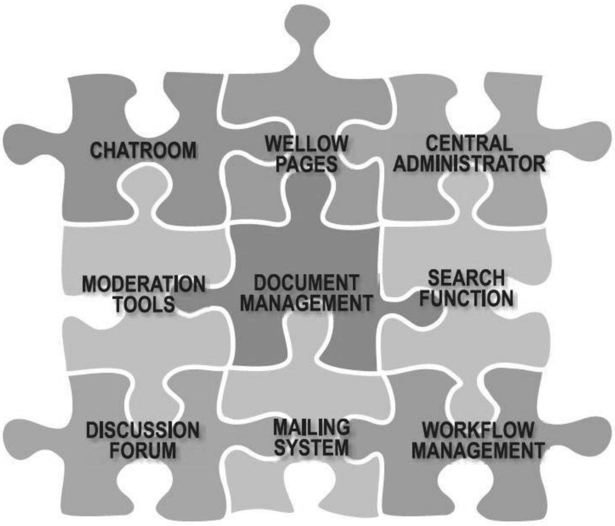


Fig. 2. Components of the CoP platform [5]

The technology for the development of the web platform will be determined during its implementation. However, it should be taken into account the need for a dynamic website able to ensure users interactions and to facilitate content update by the users. The minimum set of technologies that could be used are HTML5, CSS3, jQuery. To develop a dynamic part .NET Framework, ASP.NET, C#/Visual Basic will be used. The database that will serve the dynamic part will be managed by MS SQL Server. The versions of the server and the database server technology will be determined by the company hosting the website.

In order to fine tune the concept of the NABOO web platform, it is essential to conduct a survey among NABOO members in order to understand their requirements for the services to be offered, as well as to investigate the structure of the web sites of similar organizations. For example, a brief look on the platforms of the European Academy of Optometry and Optics (EAOO) and that of the European Council of Optometry and Optics (ECOO) shows that the available functionalities are to a large extend similar to those envisaged in the concept for the NABOO web platform. Interesting are the ideas for the formation of groups of interest in the professional sphere as a type of CoP. Another success factor is to carefully select a suitable IT partner for the design and maintenance of the web platform, and afterwards choosing the most appropriate technologies meeting the concept goals.

4 Conclusion

The project for design of a web platform for knowledge management at NABOO will be implemented step-by-step. First, it is decided that the web platform will be designed on a fully new concept in a domain owned by NABOO, and with the assistance of software developers from Sofia University, FMI. By a coincidence, the development of the web platform of NABOO has started with a specific part, e.g. supporting the organization of the ECOO General Assembly to be hosted by NABOO in Sofia in October 2013. The further development will focus on the presentation of the organization, its members and a news section. In parallel, it is envisaged online registration of the Assembly participants, as well as inclusion of online services for NABOO members. At a later stage, it will be developed a web information system with limited access only for NABOO members, which will contain specific information and a library of texts and documents on professional topics. The presence of a discussion forum and a system for e-learning would help significantly to organizational learning and management of both individual and organizational knowledge of NABOO.

As main expected benefits could be listed the improved transfer of know-how and best practices among members, the fast dissemination of information about sectoral trends, as well as international developments of interest to NABOO community. In order to be successful, the platform should fully respect the interests and the requirement of the community. A proper training for its active use should be ensured before launching it. Therefore, the platform concept and functionalities will be widely discussed among NABOO members, and the platform prototype will be validated by a selected focus group during the testing phase.

Acknowledgements

The author wish to kindly acknowledge the support and information provided by NABOO management for this paper and the financial support ensured by the European Social Fund through the Human Resource Development Operational Programme under contract BG051PO001-3.3.06-0052 (2012/2014).

References

1. Gourova, E., Antonova, A., Nikolov, R. (eds.): Knowledge Management. Bulvest 2000, Sofia. (2012) ISBN 978-954-18-0839-9 (in Bulgarian)
2. Hafkesbrink, J., Evers, J.: Innovation 3.0: Embedding into community knowledge – The relevance of trust as enabling factor for collaborative organizational learning. In: Hafkesbrink, J. et al. (eds.) Competence Management for Open Innovation, Josef EUL Verlag, Köln, Germany. (2010)

3. Hildreth, P., Kimble, C.: Knowledge Networks: Innovation through Communities of Practice. Idea Group Publishing, Hershey. (2004)
4. `http://www.naboo-bg.com`
5. TRAINMOR KNOWMORE consortium: Handbook on Organizational Knowledge Management in European Organizations. Sofia (2008)
6. Schubert, P., Wölfle, R.: The eXperience Methodology for Writing IS Case Studies. In: Proceedings of the Thirteenth Americas Conference on Information Systems (AMCIS). (2007)

Author Index