DOI: 10.22616/REEP.2018.001

An Innovative Method for Data Mining in Higher Education

Andreas Ahrens[1] Dr. ing. habil.; **Jeļena Zascerinska**[2] Dr. paed.
Julija Melnikova[3] Dr. sc.soc.; **Natalia Andreeva**[4] Dr. paed.
Hochschule Wismar, Germany[1]; Centre for Education and Innovation Research, Latvia[2]
Klaipeda University, Lithuania[3]; Immanuel Kant Baltic Federal University, Russia[4]
andreas.ahrens@hs-wismar.de[1]; knezna@inbox.lv[2]
julija.melnikova@ku.lt[3]; andreeva_natalia@list.ru[4]

Abstract: Efficiency of process remains the key issue in higher education. Process efficiency is closely connected with data mining as data mining supports decision making in higher education. Development of Information and Communication Technology (ICT) has promoted the emergence of large data sets or, in other words, big data in all the areas of higher education. The aims of the research are to analyse scientific literature on innovative methods for data mining in higher education as well as to highlight advantages of the innovative method for data mining in higher education through the comparison with other methods for data mining. The methodology of the present research is built on the inter-related steps following a logical chain: analysis of scientific literature on innovative methods for data mining in higher education → comparison of innovative methods for data mining in higher education with other methods of data mining → advantages of the innovative method for data mining in higher education → conclusions. Exploratory research was employed in the present investigation. Exploratory research is aimed at generating new research questions. Interpretive paradigm was applied to the analysis. The analysis of scientific literature reveals the theoretical inter-connections between data analysis, data analytics, data mining, burstiness and gap processes. Burst detection method based on gap processes is identified as an innovative method for data mining in higher education. Such advantages of the innovative method, namely burst detection method base on gap processes, for data mining in higher education are disclosed: a realistic evaluation of burstiness in a process, and a given precision in analysing burstiness parameters/variables such as probability and concentration. Application of the burst detection method base on gap processes for data mining in higher education supports decision making for increasing efficiency in such processes of higher education as predicting student performance, planning and scheduling, enrolment management, target marketing, management and generation of strategic information, students' selection of courses, measurement of students' retention rate, grant fund management of an institution, optimization of study processes. Directions of further research are proposed.

Keywords: higher education, big data, data analytics, data mining, burst detection.

Introduction

Efficiency of process remains the key issue in many if not in all the life fields such as business, industry, medicine, and education that includes higher education, too. Process efficiency is closely connected with data mining as data mining supports decision making in a variety of processes including processes in higher education.

Development of Information and Communication Technologies (ICT) has promoted the emergence of large data sets or, in other words, big data in all the areas of our life including higher education as shown in Figure 1.

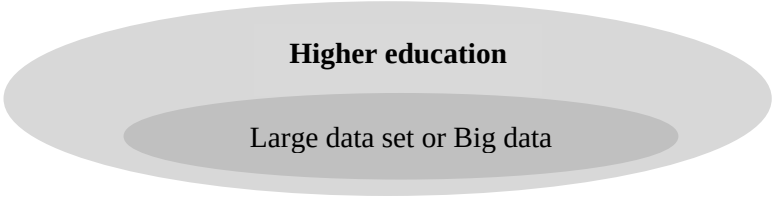

Figure 1. The relationship between higher education and large data sets.

The proliferation, ubiquity and increasing power of computer technology has dramatically increased data collection, storage, and manipulation ability. Data sets have grown in size and complexity, too (Dermino, Fortingo, 2015).

Big data serves as support in decision making for increasing efficiency of higher education in such areas as data analysis and visualisation, predicting student performance, student observation, educator observation, student grouping, planning and scheduling, enrolment management, target marketing, management and generation of strategic information, students' life cycle management, students' selection of courses, measurement of students' retention rate and the grant fund management of an institution (Goyal, Vohra, 2012, 116-119), re-structuring higher education institution, optimization of study processes (Ahrens et al., 2016a), and reforming of higher education system.

In order to enable holistic contextual decisions in higher education, data analytics and data analysis have to be carried out. It should be noted that data analysis is considered to be a human being activity or, in other words, direct "hands-on". In turn, data analytics is identified as a mechanical or algorithmic process or, in other words, indirect, automated data processing (Dermino, Fortingo, 2015). Data analytics in general and big data analytics in particular driven by data mining (Apte, 2011). Figure 2 reveals the relationship between data analysis, data analytics and data mining.

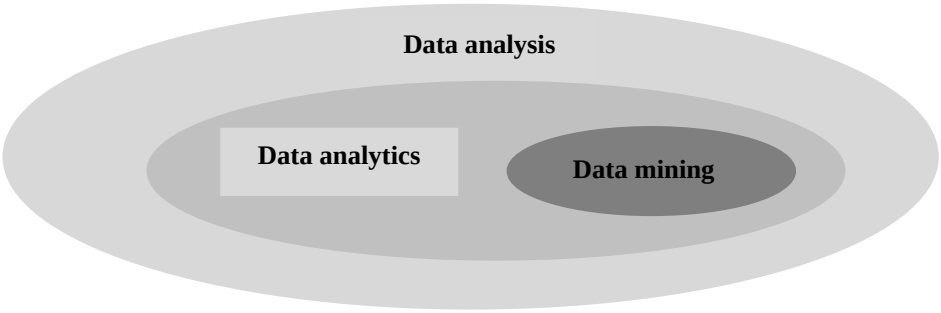

Figure 2. The relationship between data analysis, data analytics and data mining.

Data mining serves as a tool:

- on the one hand, to designate the tendency in the field of investigation (Pierrehumbert, 2012),
- on the other hand, to assist in discovering new patterns from large data sets according to different perspectives for categorization.

Data mining or, in other words, automated detection has been already aided by other discoveries in computer science, such as neural networks, cluster analysis, genetic algorithms (1950s), decision trees and decision rules (1960s), and support vector machines (1990s) (Dermino, Fortingo, 2015). However, effective methods and approaches for automated detection are still an open research area that is constantly being developed.

The aims of the research are to analyse scientific literature on innovative methods for data mining in higher education as well as to highlight advantages of the innovative method for data mining in higher education through the comparison with other methods of data mining underpinning elaboration of a new research question on efficient application of the innovative method for data mining in higher education.

The guiding ***research questions*** are as following:

- what is an innovative method for data mining in higher education;
- what are advantages of an innovative method for data mining in higher education?

Methodology

The methodology of the present research is built on the inter-related steps following a logical chain: analysis of scientific literature on innovative methods for data mining in higher education → comparison of the innovative method for data mining in higher education with other methods for data mining → outline of the advantages of the innovative method for data mining in higher education → conclusions. Exploratory research was employed in the present investigation. Exploratory research is aimed at generating new research questions (Phillips, 2006).

The exploratory methodology proceeds (Melnikova, Zascerinska, Glonina, 2015):

- from exploration in Phase 1;
- through analysis in Phase 2;
- to generating a new research question in Phase 3.

The interpretive paradigm was applied to the investigation. The interpretive paradigm aims to understand other cultures, or, in other words, other scientific disciplines, and establishment of ethically sound relationships (Taylor, Medina, 2011). Interpretative paradigm is characterized by the researcher's practical interest in the research question (Cohen, Manion, Morrsion, 2005). The researcher is the interpreter.

Results and Discussion

Analysis of scientific literature allows drawing up a conclusion that burst detection method as a method for automated detection has recently attracted a lot of research interests in a variety of research fields such as text stream (Kalogeratos, Zagorisios, Likas, 2016), reviews for review spammer detection (Fei et al., 2013), e-business (Ahrens, Zascerinska, 2017). Intense research activities on burst patterns were carried out (Subasic, Berendt, 2010). In higher education, the research efforts in applying burst detection method focused on practical use of the model for simulation of study process optimization in rural areas (Ahrens et al., 2016a). The model was built on the bursty nature of students (Ahrens et al., 2016a). Burst means sudden concentration in time periods (Fei et al., 2013). The burst detection method exploits the bursty nature of a phenomenon (Fei et al., 2013). Phenomenon's burstiness is revealed as phenomenon's frequency at an unusual high rate (Kalogeratos, Zagorisios, Likas, 2016).

Interval of high-activity alternating with long low-activity periods can be found in many areas of our daily life. Table 1 reflects the phenomenon of burstiness in a range of scientific fields (Ahrens, Zascerinska, 2016) as part of our daily life.

Table 1

Burstiness in different scientific fields

Scientific field	Phenomenon of burstiness
Telecommunications	Burstiness of bit-errors in data transmission.
Economics	Burstiness of crises.
Natural sciences	Burstiness of disasters, earthquakes.
Logistics	Burstiness of traffic.
Social media	Burstiness of hot topic, keyword, event.
Business	Burstiness of workload.
E-Business	Burstiness of buyers.

A classic example is the distribution of bit-errors in telecommunication systems. Here, intervals with low bit-errors are surrounded by intervals with high number of bit-errors. Beginning in the 1960s E.N. Gilbert presented the first model in telecommunications which emphasized that bit errors occurred in bundles or, in other words, bursts (Gilbert, 1960; Elliott, 1963). This work has later been extended by H. Wilhelm who introduced some closed form solutions for describing the bit-error distributions in wireless communication channels such as the short-wave transmission channel (Wilhelm, 1976) by introducing regenerative model approaches. These investigations were encouraged by practical measurement campaigns in the sixties and seventies. H. Wilhelm introduced already at that time simulation models such as the L-model or the A-model which took the effect of burstiness into consideration. He recognized that the bit error probability (also sometimes referred as bit error rate) is not sufficient to describe the effect of burstiness in wireless communication. Instead he defined solutions which take burstiness into account by defining models with two input parameters such as the bit error rate and the error concentration value.

H. Wilhelm mapped the process of bit errors in telecommunication systems onto processes defined by gaps between two consecutive bit errors. Since the gap-length undergoes some variations, the statistical description requires appropriate probability distribution functions. By defining a gap-distribution function (defined as the probability that a gap between two bits in larger than k bits) or a gap-density function (defined as the probability that a gap between two bits equals k bits) he could find some closed form solutions. The model characteristic has later been extended by A. Ahrens (Ahrens, 2000). Supported by practical measurements, these models make use of the assumption that the block error rate (i.e. a block with at least one-bit error) can be described as a function of the bit-error probability and the

block length. In the double-logarithmic scale the linearity between the block error rate and the block length is used to define the simulation model characteristic as well as is used to define the inherent concentration between consecutive bit-errors. The model characteristic is proved by many measurements campaigns.

Digital simulation models such as the beforehand mentioned models for describing burstiness in wireless transmission systems are an important prerequisite for optimizing the underlying components for data transmission such as transmitting or receiving algorithms. Such simulation models have been heavily used for optimizing of coding schemes. So was the probability of undetected errors for shortened Hamming codes investigated by C. Lange and A. Ahrens (Lange, Ahrens, 2001) on bursty channels. Another example showing the importance of such simulation models is the modelling of connection arrivals in Ethernet-based data networks (Kessler et al., 2003), where the intervals between consecutive data packets were analysed in a data network.

Thus, burst detection method based on gap processes is identified as an innovative method for data mining. Table 2 demonstrates models or, in other words, methods, for evaluation of burstiness in a number of scientific fields.

Table 2

Comparison of models for evaluation of burstiness in a number of scientific disciplines

Model's element	**Scientific fields**			
	Social media	**Reviews**	**Text stream**	**Higher education**
Criteria	Burstiness of hot topic, keyword in a sequence of batched georeferenced documents	Reviewers' co-occurrence in bursts	Term burstiness and co-burstiness	Students' burstiness
Indicators	Locality	Smoothness	Frequency	Students' probability
		Continuity		Students' concentration
Feature	Sequence of batched geo-referenced documents	Individual reviewer behavioural features	Consecutive batches of documents	Sequential independence of gaps between two students or sequentially independent gaps of length *k* between the individual students
Methodo-logical back-ground	J. Kleinberg's burst detection algorithm, which is based on a queuing theory for detecting bursty network traffic	Kernel Density Estimation (KDE) techniques	Two-state automaton by J. Kleinberg (Kleinberg, 2003)	Gap distribution function within a sequence of the disturbed and interrupted transmission intervals

For comparison purposes, the criterion and indicator of analysis of burstiness of hot topic, keyword, event in a sequence of batched georeferenced documents in social media is developed by a group of Japanese researchers as geo-annotated user-generated data on social media sites is becoming one of the most influential sources of information (Kotozaki, Tamura, Kitakami, 2015). This group of Japanese researchers built their model of evaluation of burstiness of hot topic, keyword in a sequence of batched georeferenced documents on Kleinberg's burst detection algorithm, which is based on a queuing theory for detecting bursty network traffic (Kotozaki, Tamura, Kitakami, 2015).

Another research group involved researchers from the USA. The research group proposed such a method as relationships among reviewers by linking reviewers in a burst (Fei et al., 2013, 176). The method focuses on individual reviewer behavioural features. Such properties or criteria as smoothness and

continuity are desirable properties for review burst detection in a product (Fei et al., 2013, 176). The method for burst detection employs Kernel Density Estimation (KDE) techniques.

An international research group from France and Greece emphasized term burstiness and co-burtsiness for the improvement of text streams clustering (Kalogeratos, Zagorisios, Likas, 2016). Frequency is considered as the indicator within consecutive batches of documents. The methodological background of the method for the improvement of text streams clustering is based on two-state automaton by Kleinberg (Kleinberg, 2003).

The comparative analysis, reflected in Table 1, of the methods, namely the model of evaluation of burstiness of hot topic, keyword in social media shown by the group of Japanese researchers (Kotozaki, Tamura, Kitakami, 2015), reviewers' co-occurrence in bursts revealed by the of researchers from the USA (Fei et al., 2013), term burstiness and co-bursiness disclosed by the international research group from France and Greece (Kalogeratos, Zagorisios, Likas, 2016) and the model for evaluation of students' burstiness in study process (Ahrens et al., 2016a), was grounded on the comparison of the models' elements such as:

- criteria,
- indicators,
- feature,
- methodological background.

The comparative analysis allows identifying burst detection method based on gap processes to be an innovative model or, in other words, method for data mining in higher education. The model is mathematical, namely ($Y \equiv u(k)$ (Ahrens et al., 2016a). The model is applicable within the binary option paradigm: "to be, or not to be" formulated already in 1603 by William Shakespeare in his play *Hamlet* (Shakespeare, 1603). The model is based on gap processes. Gap in the present contribution means a process ends without an outcome (Ahrens et al., 2015). The analysis assists in concluding that a process is characterized by sequential independence of gaps between two phenomena (Ahrens, Zascerinska, 2016). Criteria for process optimization were defined as probability and concentration (Ahrens et al., 2016a). Further on, the levels of burstiness (Ahrens et al., 2016b) are summarized in Table 3.

Table 3

Levels of burstiness

Characteristics	**Levels**				
	L1	**L2**	**L3**	**L4**	**L5**
	Very low	**Low**	**Average**	**High**	**Very high**
Probability p_e	0.0 – 0.10	0.11 – 0.39	0.41 – 0.59	0.60 – 0.79	0.80 – 1.0
Concentration $(1-\alpha)$					

In order to outline advantages of the innovative method for data mining, a comparative analysis of other methods that take phenomenon's burstiness into account is to be carried out. However, many researchers highlight that there is a lack of common procedures that makes it impossible to compare methods in a principled way (Subasic, Berendt, 2010).

By advantages, any trait, feature or aspect that gives an individual, entity or any other thing a more favourable opportunity for success are identified (Melnikova et al., 2017). Advantages are outlined through structuring and summarizing content analysis (Mayring, 2014).

Structuring and summarizing content analysis (Mayring, 2014) allows highlighting such advantages of the innovative method, namely burst detection method based on gap processes, for data mining in higher education:

- a realistic evaluation of burstiness in study process;
- a given precision in analysing burstiness parameters/variables such as:
  - probability,
  - concentration.

Conclusions

The analysis of scientific literature assists in revealing the theoretical inter-connections between data analysis, data analytics, data mining, burstiness and gap processes. In scientific literature burst detection method is identified as an innovative method for data mining in general and in higher education in particular. The innovative method, namely burst detection method, for data mining in higher education is based on gap processes. Phenomenon's probability and concentration are the indicators for analysing burstiness in a variety of processes including processes in higher education. Structuring and summarizing content analysis within the comparative analysis of burst detection methods applied to social media, reviews, text stream and higher education facilitated the outline of such advantages of the innovative method, namely burst detection method based on gap processes, for data mining in higher education as:

- a realistic evaluation of burstiness in a process;
- a given precision in analysing burstiness parameters/variables such as:
  - probability,
  - concentration.

Application of the innovative method, namely burst detection method based on gap processes, for data mining in higher education supports decision making for increasing efficiency of higher education in such areas as data analysis and visualisation, predicting student performance, student observation, educator observation, student grouping, planning and scheduling, enrolment management, target marketing, management and generation of strategic information, students' life cycle management, students' selection of courses, measurement of students' retention rate and the grant fund management of an institution, re-structuring higher education institution, optimization of study processes, and reforming of higher education system.

The new research question is put forward: What are conditions for efficient application of the innovative method for data mining in higher education?

The present research has limitations. The inter-connections between data analysis, data analytics, data mining and burstiness have been set. The study is also limited by the implementation of the analysis of scientific literature only. Another limitation is only a few publications on methods for data mining in higher education. Therein, the results of the research cannot be representative for the whole area. Nevertheless, the results of the research, namely the advantages of the innovative method for data mining in higher education, may be used as a basis of analysis of a variety of processes in higher education such as study process, management process and teaching process. If the results of other researches had been available for analysis, different results could have been attained. There is a possibility to continue the study.

Further research tends to facilitate the practical applications of the innovative method, namely burst detection method based on gap processes, for data mining in higher education. Interdisciplinary research could enhance relevant tools and techniques of the innovative method, namely burst detection method based on gap processes, for data mining in higher education. Analysis of disadvantages of the innovative method, namely burst detection method based on gap processes, for data mining in higher education is proposed, too. Further research tends to disclose recommendations on practical application of the innovative method, namely burst detection method based on gap processes, for data mining in higher education. A comparative research on methods as well as their tools and techniques for data mining in higher education could be carried out, too.

Bibliography

1. Ahrens A. (2000). A new digital radio-channel model suitable for the evaluation and simulation of channel effects. In *Speech Coding for Algorithms for Radio Channels*. London: IET.
2. Ahrens A., Purvinis O., Zascerinska J., Andreeva N. (2015). Gap Processes for Modelling Binary Customer Behavior. In N. Grünwald, M. Heinrichs (Eds.), Proceedings of the International Conference on *Engineering and Business Education*, 8. Fredrikstad: Ostfold University College and University of Wismar, 8-13.
3. Ahrens A., Purvinis O., Zascerinska J., Andreeva N. (2016a). A Model for Simulation of Study Process Optimization in Rural Areas. In V. Dislere (Ed.), Proceedings of the International

Scientific Conference *Rural Environment, Education, Personality (REEP)*, 9. Jelgava: LLU, 145-152. Retrieved from: http://llufb.llu.lv/conference/REEP/2016/Latvia-Univ-Agricult-REEP-2016proceed2255-808X-145-152.pdf.

4. Ahrens A., Purvinis O., Zascerinska J., Andreeva N. (2016b). Education for Health Society: Indicators of Burstiness in Research. *Education in a Changing Society*, 1, 28-40. Retrieved from http://journals.ku.lt/index.php/educs/article/download/1315/1714
5. Ahrens A., Zascerinska J. (2016). Gap Processes for Analysing Buyers' Burstiness in E-Business Process. In Ch. Callegari, M. van Sinderen, P. Sarigiannidis, P. Samarati, E. Cabello, P. Lorenz, M. S. Obaidat (Eds.), Proceedings of the 13th International Joint Conference on *e-Business and Telecommunications (ICETE 2016)*, 2 (ICE-B). Lisbon: Science and Technology Publications Lda, 78-85.
6. Ahrens A., Zascerinska J. (2017). E-Shop Visitors' Burstiness as a Predictor of Performance – The Case of eBay. In M. van Sinderen, M.S. Obaidat, E.Cabello (Eds.), Proceedings of the 14th International Joint Conference on *e-Business and Telecommunications (ICETE 2017)*, 4 (ICE-B). Madrid: Science and Technology Publications, Lda, 78-82. Retrieved from http://www.scitepress.org/DigitalLibrary/PublicationsDetail.aspx?ID=7krlu636A/M=&t=1
7. Apte C. (2011). *The Role of Data Mining in Business Optimization*. IBM Corporation. Retrieved from https://www.siam.org/meetings/sdm11/apte.pdf
8. Cohen L., Manion L., Morrsion K. (2005). *Research Methods in Education*. (5th ed.). London and New York: Routledge/Falmer Taylor and Francis Group. Retrieved form https://research-srttu.wikispaces.com/file/view/Research+Methods+in+Education_ertu.pdf
9. Dermino F., Fortingo K. (2015). What is Data Mining Methods with Different Group of Clustering and Classification. *American Journal of Mobile Systems, Applications and Services*, 1(2), 140-151. Retrieved from http://files.aiscience.org/journal/article/pdf/70110028.pdf
10. Elliott E.O. (1963). Estimates of Error Rates for Codes on Burst-Noise Channels. *Bell System Technical Journal*, 42(5), 1977–1997.
11. Fei G., Mukherjee A., Liu B., Hsu M., Castellanos M., Ghosh R. (2013). Exploiting Burstiness in Reviews for Review Spammer Detection. In E. Kiciman, N. B. Ellison, B. Hogan, P. Resnick, I. Soboroff (Eds.), Proceedings of the International AAAI Conference on Weblogs and Social Media, *ICWSM*, 7. California: The AAAI Press. Retrieved from https://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/viewFile/6069/6356
12. Gilbert E.N. (1960). Capacity of a Burst-Noise Channel. *Bell System Technical Journal*, 39, 1253-1265.
13. Goyal M., Vohra R. (2012). Applications of Data Mining in Higher Education. *International Journal of Computer Science*, 9 (2), 113.
14. Kalogeratos A., Zagorisios P., Likas A. (2016). Improving Text Stream Clustering using Term Burstiness and Co-Burstiness. In N. Bassiliades (Ed.), Proceedings of the 9th Hellenic Conference on *Artificial Intelligence, SETN '16* (Article No. 16). Thessaloniki. Retrieved from http://kalogeratos.com/psite/files/MyPapers/CBTC-SETN2016.pdf
15. Kessler T., Ahrens A., Lange C., Melzer H.D. (2003). Modelling of connection arrivals in Ethernet-based data networks. In The 4rd International Conference on *Information, Communications and Signal Processing and 4th IEEE Pacific-Rim Conference on Multimedia (ICICS-PCM)*, 3. Singapore, 15–18. Retrieved from https://www.researchgate.net/publication/4072499_Modelling_of_connection_arrivals_in_Ethernet-based_data_networks.
16. Kleinberg J. (2003). Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery*, 7(4), 373- 397. Retrieved from https://www.cs.cornell.edu/home/kleinber/bhs.pdf
17. Kotozaki Sh., Tamura K., Kitakami H. (2015). Identifying Local Burstiness in a Sequence of Batched Georeferenced Documents. *International Journal of Electronic Commerce Studies*, 6(2), 269–288. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.873.5016&rep=rep1&type=pdf
18. Lange C., Ahrens A. (2001). On the Undetected Error Probability for Shortened Hamming Codes on Channels with Memory. In B. Honary (Ed.), IMA International Conference on *Cryptography and Coding*. Heidelberg: Springer, 9–19.
19. Mayring P. (2014). Qualitative Content Analysis: theoretical foundation, basic procedures and software solution. Klagenfurt: Social Science Open Access Repository. Retrieved from

https://www.psychopen.eu/fileadmin/user_upload/books/mayring/ssoar-2014-mayring-Qualitative_content_analysis_theoretical_foundation.pdf

20. Melnikova J., Zascerinska J., Ahrens A., Hariharan R., Clipa O., Sowinska-Milewska D., Andreeva N. (2017). A Comparative Study of Educators' Views on Advantages and Disadvantages of Open Educational Resources in Higher Education. In V. Lubkina, S. Urca, A. Zvaigzne (Eds.), Proceedings of the International Scientific Conference *Society. Integration. Education*, 1. Rezekne: Rezekne Academy of Technologies, 294-304. Retrieved form http://journals.ru.lv/index.php/SIE/article/view/2362/2305
21. Melnikova J., Zascerinska J., Glonina O. (2015). A Conceptual Framework on Entrepreneurship Education in Vocational Teachers Training. In Proceedings of the International Conference *Young Scientist*, 10. Riga: Riga Teacher Training and Educational Management Academy, 60-69.
22. Phillips D. (2006). Comparative Education: Method. *Research in Comparative and International Education*, 1 (4), 304-319.
23. Pierrehumbert J.B. (2012). Burstiness of Verbs and Derived Nouns. In D. Santos, K. Linden, W. Ng'ang'a (Eds.), *Shall we Play the Festschrift Game? Essays on the Occasion of Lauri Carlson's 60th Birthday*. Heidelberg: Springer Verlag.
24. Shakespeare W. (1603). *The Tragicall Historie of Hamlet Prince of Denmarke. The First Quarto*. London: Nicholas Ling and J. Trundell. Retrieved from http://internetshakespeare.uvic.ca/Library/facsimile/book/BL_Q1_Ham/
25. Subasic I., Berendt B. (2010). From bursty patterns to bursty facts: The effectiveness of temporal text mining for news. In Proceedings of the conference on *European Conference on Artificial Intelligence (ECAI)*, 19**.** Amsterdam: IOS Press, 517-522.
26. Taylor P.C., Medina M.N.D. (2011). Educational Research Paradigms: From Positivism to Pluralism. *College Research Journal*, 1 (1), 9-23.
27. Wilhelm H. (1976). *Datenubertragung (Data Transfer)*. Berlin: Militarverlag. (in German)