DOI: 10.17323/2587-814X.2021.1.30.46

Trends in data mining research: A two-decade review using topic analysis

Yuri A. Zelenkov
E-mail: yzelenkov@hse.ru

Ekaterina A. Anisichkina
E-mail: eaanisichkina@edu.hse.ru

National Research University Higher School of Economics
Address: 20, Myasnitskaya Street, Moscow 101000, Russia

Abstract

This work analyzes the intellectual structure of data mining as a scientific discipline. To do this, we use topic analysis (namely, latent Dirichlet allocation, LDA) applied to the proceedings of the International Conference on Data Mining (ICDM) for 2001–2019. Using this technique, we identified the nine most significant research flows. For each topic, we analyze the dynamics of its popularity (number of publications) and influence (number of citations). The central topic, which unites all other direction, is General Learning, which includes machine learning algorithms. About 20% of the research efforts were spent on the development of this direction for the entire time under review, however, its influence has declined most recently. The analysis also showed that attention to topics such as Pattern Mining (detecting associations) and Segmentation (object separation algorithms such as clustering) is decreasing. At the same time, the popularity of research related to Recommender Systems, Network Analysis, and Human Behaviour Analysis is growing, which is most likely due to the increasing availability of data and the practical value of these topics. The research direction related to practical Applications of data mining also shows a tendency to grow. The last two topics, Text Mining and Data Streams have attracted steady interest from researchers. The results presented here shed light on the structure and trends of data mining over the past twenty years and allow us to expand our understanding of this scientific discipline. We can argue that in the last five years a new research agenda has been formed, which is characterized by a shift in interest from algorithms to practical applications that affect all aspects of human activity.

Key words: data mining topics, topic analysis, scientometrics.

Citation: Zelenkov Yu.A., Anisichkina E.A. (2021) Trends in data mining research: A two-decade review using topic analysis. *Business Informatics*, vol. 15, no 1, pp. 30–46. DOI: 10.17323/2587-814X.2021.1.30.46

Introduction

The term "data mining" (DM) appeared in the 1960s to describe the search for correlations without an a priori hypothesis [1]. According to the widely accepted definition that is used in many textbooks now, data mining is the extraction of implicit, previously unknown, and potentially useful information from data [2, 3]. Besides, Rather [4] defines data mining as a combination of three easy concepts:

✦ statistics that include the classical descriptive tools, e.g. degrees of freedom, *F*-ratios, and *p*-values, but exclude inferential conclusions;

✦ big data as an umbrella term for datasets of any size with the accent on big size since a tremendous amount of data impacts almost every aspect of our lives;

✦ machine learning (ML), i.e. tools to build computer programs that sift through databases automatically, seeking regularities or patterns [2].

Statistics and machine learning provide the technical basis of data mining. They are used to extract information from the raw data. Some authors also view DM as part of the process for knowledge discovery from data (KDD). This process may include techniques such as data preprocessing (cleaning and integration), data storage, online analytical processing, data cubes, etc. [3].

As follows from these definitions, data mining is a scientific discipline that combines achievements in several areas of research. The structure of any scientific discipline can be represented as a set of evolving topics, i.e. significant, implicit associations hidden in fragmented knowledge areas. Trends in these topics (for example, a change in the number of publications and their citation) reflect a shift in the interests of the research community. In particular, the study of this dynamic allows us to determine the most relevant areas of research in the present and extrapolate them in the near future. In addition, understanding the fundamental shifts in the interests of researchers helps us to determine the place of the studied discipline in the general body of human knowledge, its interaction with other disciplines and the overall contribution to human progress.

The traditional method of studying the structure of a scientific discipline is survey or review. However, due to the interdisciplinary nature of data mining, there are practically no reviews considering DM as a single discipline (yet it should be noted that surveys of narrower topics are published continuously).

In a review paper published in 2006, Yang and Wu [5] noted that data mining had achieved tremendous success. However, there is still a lack of timely exchange of essential topics in the community as a whole. Authors of [5] ranked the ten most important problems in DM:

✦ developing a unifying theory of data mining;

✦ scaling up for high dimensional data and high-speed data streams;

✦ mining sequential data and time series data;

✦ mining complex knowledge from complex data;

✦ data mining in graph-structured data;

✦ distributed data mining and mining multi-agent data;

✦ data mining for biological and environmental problems;

✦ data mining process-related problems;

✦ security, privacy, and data integrity;

✦ dealing with non-static, unbalanced and cost-sensitive data.

These problems divide the overall DM research flow into smaller, more focused segments. In 2010, Wu provided additional comments on these challenging issues [6], and they were the subject of discussion in a special panel at the 10th International Conference on Data Mining (ICDM).

Yang and Wu [5] view the development of a unified theory of DM as the most critical issue. It should be a theoretical framework that unifies different techniques designed for individual problems, including clustering, classification, association rules, etc., as well as different data mining technologies (such as statistics, machine learning, database systems, etc.). It should help the field and provide a basis for future research.

Most of the identified problems relate to algorithms for working with data types that became relevant in the 2000s (ultra-high dimensional data, high-speed data streams, time series, networks and other complex data). The authors of [5, 6] consider ecological and environmental informatics as the most important area of DM applications.

In addition to the analysis of critical challenges, work [6] presents a list of the most important topics of data mining (*Table 1*). This list was obtained based on expert opinions; hence, it can serve as a reference to the structure of the scientific discipline. However, the expert-based approach does not provide quantitative metrics that measure the relative importance of various topics and their change over time.

Liao et al. [7] presented a review of the literature on data mining techniques and applications from January 2000 to August 2011. They selected 216 articles from 159 academic journals using keywords like 'data mining,' 'decision tree,' 'artificial neural network,' 'clustering,' etc. Based on papers selected, they identified nine categories of DM techniques (systems optimization, knowledge-based systems, modeling, algorithm architecture, neural networks, etc.). In addition, the authors of [7] presented the essential trends in data mining. According to the results presented, the most important trend is the Association Rules (rank 5), followed by Neural Networks (rank 4) and then Classification and Support Vector Machines (both have rank 3). The authors do not describe the method for ranking; however, we can assume that it is based on counting the number of references on each technique in the analyzed corpus of publication.

Table 1.

Top 10 data mining topics [6]

No	Topic
1	Classification (including C4.5, CART, kNN, and Naive Bayes)
2	Statistical learning (SVM and mixture models)
3	Association analysis
4	Link mining (e.g. PageRank algorithm)
5	Clustering
6	Bagging and boosting
7	Sequential patterns
8	Integrated mining (e.g. integrating classification and association rule mining)
9	Rough sets
10	Graph mining

To the best of our knowledge, the publications cited above are the only ones that examine the dynamics of data mining as a single scientific discipline. As already noted, they are based on subjective assessments.

The idea of our work is to apply formal methods of topic analysis to publications in the field of data mining. As an object of analysis, we use the proceedings of the International Conference on Data Mining (ICDM), which has been held annually since 2001.

1. Data

The International Conference on Data Mining (ICDM) is a top conference that, along with the SIGKDD Conference on Knowledge

Discovery and Data Mining (KDD), ACM International Conference on Web Search and Data Mining (WSDM) and a few others, forms a network of major forums in the field of data mining and knowledge discovery from data. The Web of Science (WoS) database contains information on 5120 publications of the main ICDM tracks and related workshops. *Figure 1* represents the time distribution of these publications.

The WoS database contains such data as the authors, title of publication, abstract, and the number of citations that are necessary for our study.

2. Research method

One of the most popular techniques of bibliometric networks analysis is term-level coupling implemented in VOSviewer software [8]. This approach allows us to identify clusters of terms that can be viewed as more or less stable implicit structures shaping the scientific discipline. Authors of a review of literature-based discovery [9] list the main computation techniques that automate the knowledge discovery process. They noted that topic modeling that allows observing how topic-level information is propagated among documents provides more deep insight of document corpora than term-level analysis. However, topic modeling is still relatively rarely used in literature analysis [9, 10].

Mann et al. [11] used a combination of topic modeling and citation analysis to estimate the impact factor of the topic over time and topical diversity of documents in computer science. Dam and Ghose [12] used topic modeling to analyze the content of the proceedings of the International Conference on Principles and Practices of Multi-Agent Systems (PRIMA). Among recent works, Zelenkov [10] applied topic analysis to the knowledge management area. The last paper pays special attention to the topic dynamics, i.e. how a number of publications and citations regarding each topic

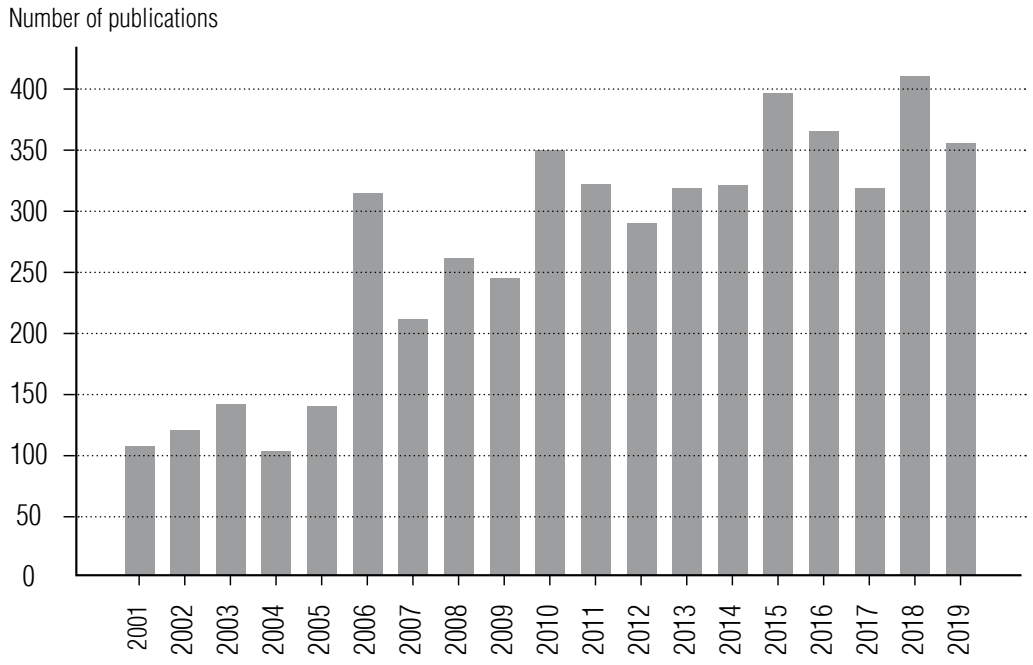

Fig. 1. Distribution of ICDM publications

changes in time. It helps shed light on the shift in research interest and identify critical trends of the present time.

Yet another application of topic analysis is presented in [13], where it is used to quantify the similarity and evolution of scientific disciplines. In [14] the authors propose topic evolution trees generated from the heterogeneous bibliographic network.

A topic is a set of words that often co-occur in texts related to a given subject area. Probabilistic topic modeling is based upon the idea that documents are mixtures of topics, where a topic is a probability distribution over terms.

Let there be a finite set of topics T, which is not known. Each use of the term w in document d is associated with some topic $t \in T$. Thus, a collection of documents is considered as a set of triples (d, w, t) selected randomly and independently from the distribution defined on a finite set $D \times W \times T$. Documents $d \in D$ and the terms $w \in W$ are observable variables. The topics $t \in T$ are latent variables that must be defined.

The topic model automatically detects latent topics by the observed frequencies of words in the documents:

$$p(w|d) = \sum_{t \in T} p(t|d)\, p(w|t).$$

Thus, the input of the algorithm is a matrix $D \times W$, which cells contain counts of the word w in document d.

To prepare matrix $D \times W$, we used abstracts of 5120 papers downloaded from the Web of Science database, as described in the previous section. According to [15], differences between abstract and full-text data are more apparent within small document collections. Therefore, we have selected abstracts as an object of analysis.

According to the general text mining technique, abstracts were tokenized, and the terms obtained were converted to standard form. Next, words that belong to an extended stop word list were deleted. The extended stop-word list includes standard English stop-words and corpus-specific words that appear in less than 5% and more than 60% of documents. We also created bigrams to join terms often co-occurred beside. As a result, we got a sparse matrix $D \times W$ with dimensions of 5120 × 1000, only 1.62% of the cells of which contain values greater than zero.

To compute the topics, we used a latent Dirichlet allocation (LDA) algorithm that is based on the additional assumption that the distribution Θ of documents θ_d and distribution Φ of topics φ_t are spawn by Dirichlet distributions [16]. To build the model, one should define a number of topics $|T|$; the LDA algorithm computes distributions Θ and Φ. As a result, each topic is presented by the weighted list of words; the weight of a word corresponds to its importance in the topic definition. The weighted list of topics presents each document; the weight of the topic corresponds to its significance in the document.

Determining the number of topics is a critical issue in topic analysis; many authors use various kinds of grid search optimizing a specific metric [10]. We used more advanced techniques, namely, Bayesian optimization [17]. Such an approach allows us to optimize simultaneously not only the number of the topics and also parameters of Θ and Φ distributions and other parameters of the algorithm. The optimization target is a perplexity which measures the convergence of a model with a given vocabulary W:

$$P(D) = \exp\left[-\frac{1}{N}\sum_{d \in D}\sum_{w \in d} n_{dw} \ln p(w|d)\right].$$

The perplexity of collection D is a measure of the language quality and is often used in computational linguistics. In our case, language is the distribution of words in documents $p(w|d)$. The less perplexity, the more uneven this distribution.

In general, the approach presented satisfies the guidelines to LDA users presented in [18].

An additional metric that we use to assess the quality of the model is diversity, i.e. the entropy of the distribution of words that characterize the topic:

$$H_t = -\frac{1}{\ln n_w}\sum_i^{n_w} p_t(w_i)\ln p_t(w_i), \qquad (1)$$

where n_w is the number of words describing topics;

$p_t(w_i)$ is the weight of i-th word in the topic t.

Since this metric is normalized by the number of features (words), it's possible values are in the range [0; 1]. The value 0 corresponds to the maximum focus when only one term describes the topic. Value 1 determines the situation when all the features are present in the description of the topic with the same weights, i.e. it is not identified. In a valuable model, the values of this metric should be the small and approximately the same for all topics.

When the optimal number of topics and corresponding topic distribution for each document are found, we can study topic dynamics. Let θ_{dt} is the weight of topic t in document d ($0 \le \theta_{dt} \le 1$). So, the overall popularity of topic across all documents can be defined as [10]:

$$\hat{\theta}_t = \frac{1}{|D|}\sum_{d\in D}\theta_{dt}. \qquad (2)$$

To measure the topic popularity in a particular year y it is enough to set $D = D_y$ in (2), where D_y is the set of all papers in year y.

Let C_d is the number of citations of document d and $C = \sum_{d\in D} C_d$. An impact of the topic can be defined as [10]:

$$\hat{i}_t = \frac{1}{C}\sum_{d\in D}\theta_{dt}C_d \ . \qquad (3)$$

By analogy, to obtain the topic impact in the particular year, one should set $D = D_y$ in (3).

3. Results and discussion

Performing all preprocessing operations described in the previous Section and 100 iterations of Bayesian optimization of the LDA model, we found that the optimal number of topics is 9, and the corresponding value of perplexity is 568.75.

Analyzing the dominant terms (*Figure 2*), we can conclude that each topic represents some coherent area of research. The weights of topics in documents are either large (i.e. the topic is strongly related) or near zero (i.e. the topic is unrelated).

Thus, to assign the labels, we analyzed the term distributions and most representative papers for each topic. To select the most representative papers, we sorted the publications by the topic weight and next by the number of citations, both in descending order. *Figure 2* presents the labels assigned, and *Table 2* lists the topics description.

Table 2 also presents the values of diversity, popularity and impact for each topic in the entire collection D, calculated in accordance with (1), (2) and (3), respectively. Please note that the sum of both popularity and impact is 1, so the values presented can be considered as a share of a particular topic in the total flow of data mining research, i.e. its total weight.

Additional useful information can be obtained from the analysis of the distribution of the topics' weights in the document corpora (*Figure 3*). For a more effective presentation, we excluded from the graph for each topic documents in which the weight of this topic is 0.

As follows from *Table 2*, the topic that has attracted the most attention over the past 20 years is General Learning. More than 20% of efforts in the field of data mining were spent in this direction. The works of this direction cover the widest spectrum of machine learning issues, e.g. the features selection [19] (the weight of dominant topics in this document is θ_{dt} =0.974),

Fig. 2. Visualization of the topic model using word clouds
(each word cloud represents one detected topic where the size
of words indicates the relevance of each word to that particular topic)

multi-label classification using ensembles [20] (θ_{dt} = 0.963), gradient methods [21] (θ_{dt} = 0.925), etc. These example papers were selected since they all have a large number of citations (more than 90, according to WoS). Interestingly, proceedings with the maximum weight of this topic are mainly devoted to kernel methods, e.g. [22] with θ_{dt} = 0.990. Note, that according to *Figure 3*, this topic has a weight close to 1 in the largest number of documents (more than 100). These articles focus solely on machine learning methods and do not overlap with other topics.

We defined the second important topic ($\hat{\theta}_t$ = 0.121) as Human Behavior Analysis because it focuses on the detection and prediction of patterns in the activities of groups of people and systems that these groups influence. In this area, issues are studied, such as the effect of price promotions [23] (θ_{dt} = 0.988), finding suspicious financial transactions [24] (θ_{dt} = 0.985), bitcoin volatility [25] (θ_{dt} = 0.985), and others. Note that a significant part of these works is presented at the workshops accompanying the main conference.

The next topic is Pattern Mining, as it focuses on association (rule) extraction, i.e., on the task of finding correlations between items in a dataset. Researchers study as a practical application of association rules (e.g. market basket data) and general features of patterns found in large databases. On the one hand, it can be the

Table 2.

Topics of Data Mining

Topic	Comments	Diversity	Popularity	Impact
Text Mining	Pattern detection in texts	0.779	0.107	0.110
General Learning	Machine learning algorithms and related methods like feature selection, class labeling, etc.	0.826	0.213	0.211
Segmentation	Methods based on object separation techniques: clustering, outlier detection, etc.	0.777	0.084	0.080
Applications	Practical use of data mining methods	0.826	0.097	0.095
Data Streams	Time–dependent models	0.805	0.097	0.102
Recommender systems	Algorithms that provide useful and explainable recommendations	0.799	0.076	0.079
Pattern Mining	General issues of finding correlations between items in data	0.750	0.110	0.114
Network Analysis	Community and influence flow detection in various networks	0.762	0.093	0.111
Human Behavior Analysis	Detection and prediction of patterns in the people's behavior: customer churn, market segmentation, fraud and security threats, etc.	0.844	0.121	0.096

identification of maximal frequent itemsets, i.e. an itemset that occurs in at least a systematic and realistic set of experiments [26] (θ_{dt} = 0.960). On the other hand, it can be a pattern consisting of infrequent, but highly correlated items rather than ones that occur frequently [27] (θ_{dt} = 0.987). Note that it is the most focused topic (with the lowest value of H_t).

The most representative proceedings of the Text Mining topic consider issues such as the identification and ranking of authors [28] (θ_{dt} = 0.974), topic modeling [29] (θ_{dt} = 0.969), and text clustering using semantics-based models [30] (θ_{dt} = 0.988). This is a research area with clear boundaries, which includes pattern detection in texts only and does not consider other types of unstructured data.

The research flow, which we call Data Streams, concerns time-dependent models. It includes more or less traditional analysis and prediction of time series with concept drift [31] (θ_{dt} = 0.981), and relatively more rare models, e.g. ones based on Granger causality [32] (θ_{dt} = 0.987).

The Applications topic combines works mainly devoted to the practical use of data mining methods, and which do not apply to other directions highlighted above. Examples are the detection of events using the co-locations of mobile users [33] (θ_{dt} = 0.987) and biometric security model for medical Internet of Things [34] (θ_{dt} = 0.983).

The research direction Network Analysis deals with graph models allowing us to restore the spatial structure or topology of the investi-

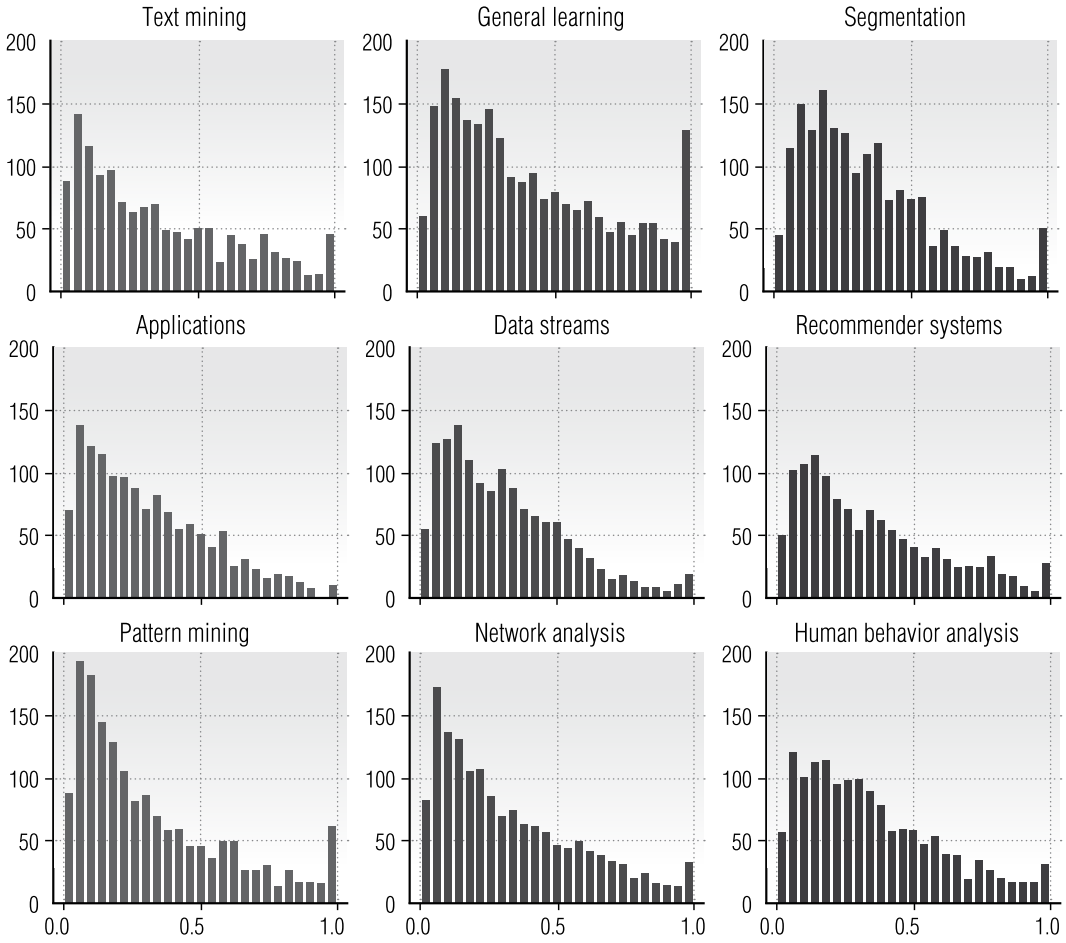

Fig. 3. Distribution of the topics' weights in the document corpora
(the horizontal axis shows the weights, the vertical axis shows the number of documents)

gated object. The most popular topic of such a kind of research is community detection using various methods of network analysis [35] (θ_{dt} = 0.983). The next issue attracting growing attention recently is the analysis of influence flows [36] (θ_{dt} = 0.985), including predictions of the popularity of messages in social networks.

We labeled the next topic as Segmentation since it includes not only a wide spectrum of clustering algorithms [37] (θ_{dt} = 0.977) but applications that are based on object separation techniques, e.g. outlier detection [38] (θ_{dt} = 0.981). Please note that according to our model, this topic is dominated in works that deal with unstructured data (images, video, sound). However, in most cases, the weight of this topic in these kinds of applications does not exceed the weights of other topics significantly.

Finally, the last but not least topic detected by our model is Recommender Systems. This direction does not need additional comments. It is only worth noting that recently researchers have paid special attention to the generation of explainable recommendations, which provides explanations about why an item is recommended [39] (θ_{dt} = 0.986).

According to *Figure 3*, in addition to General Learning, only in three other areas is there a relatively large number of articles with a topic

weight close to 1. These are Segmentation, Pattern Mining and Text Mining. These topics are also subsections of machine learning, so a relatively large number of articles focused exclusively on algorithms is published under each of these areas. On the other hand, a topic such as Applications has virtually no such focused papers. It also has a big value of diversity metric according to (1). This can be explained by the fact that works regarding practical applications, as a rule, also present new modifications of algorithms.

Our model does not distinguish artificial neural networks (ANN) as a separate direction of data mining. This contradicts [7] but is consistent with [5] and [6]. According to our results, publications that use ANN models relate most often to the areas of General Learning and Segmentation.

The next issue that should be considered is topic collaboration, i.e. the topics' co-occurrence. Let θ_{di} and θ_{dj} be the weights of the topics i and j, respectively, in document d. Thus, we can define the topics' co-occurrence in this document as a product $\theta_{di}\,\theta_{dj}$. The maximal possible value of the co-occurrence of two topics in one document is 0.25 when $\theta_{di} = \theta_{dj} = 0.5$. From this, the maximal possible value of the topics' collaboration in the document corpus is $0.25 \cdot |D|$.

Thus, the topic collaboration in the document corpus can be computed as

$$c_{ij} = \sum_{d \in D} \theta_{di} \theta_{dj}.$$

Figure 4 presents these data. General Learning can be viewed as a central topic since it is most closely related to other areas of research. Human

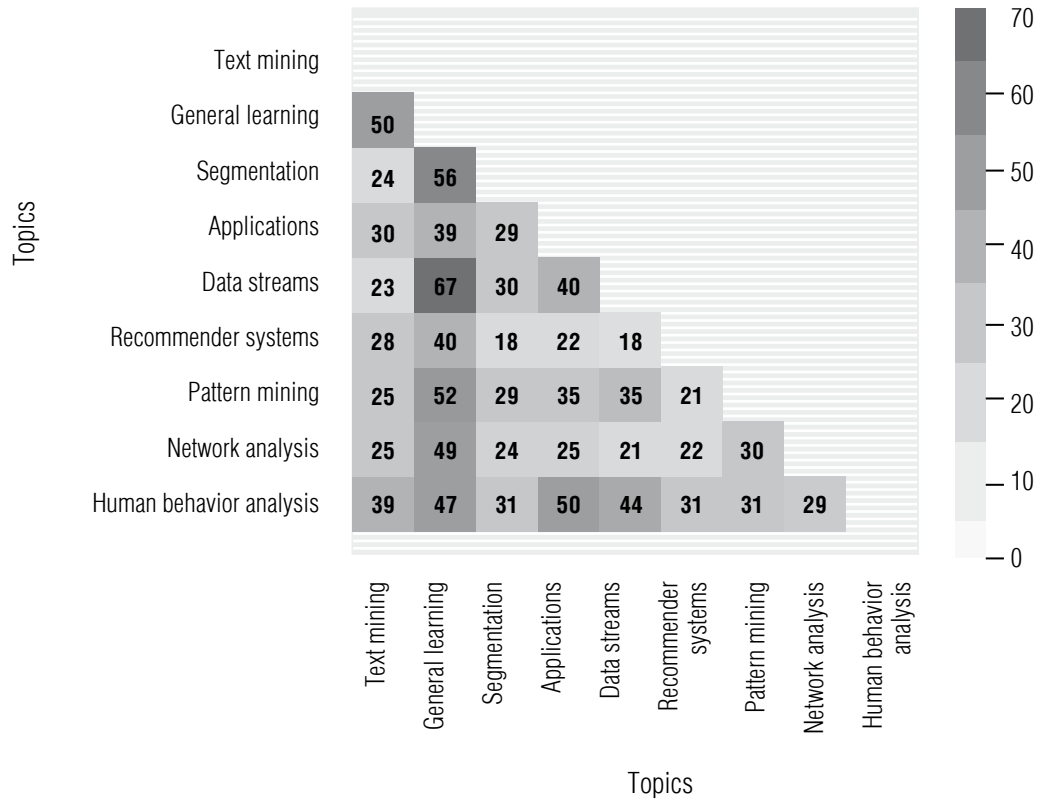

Fig. 4. A topic co–occurrence

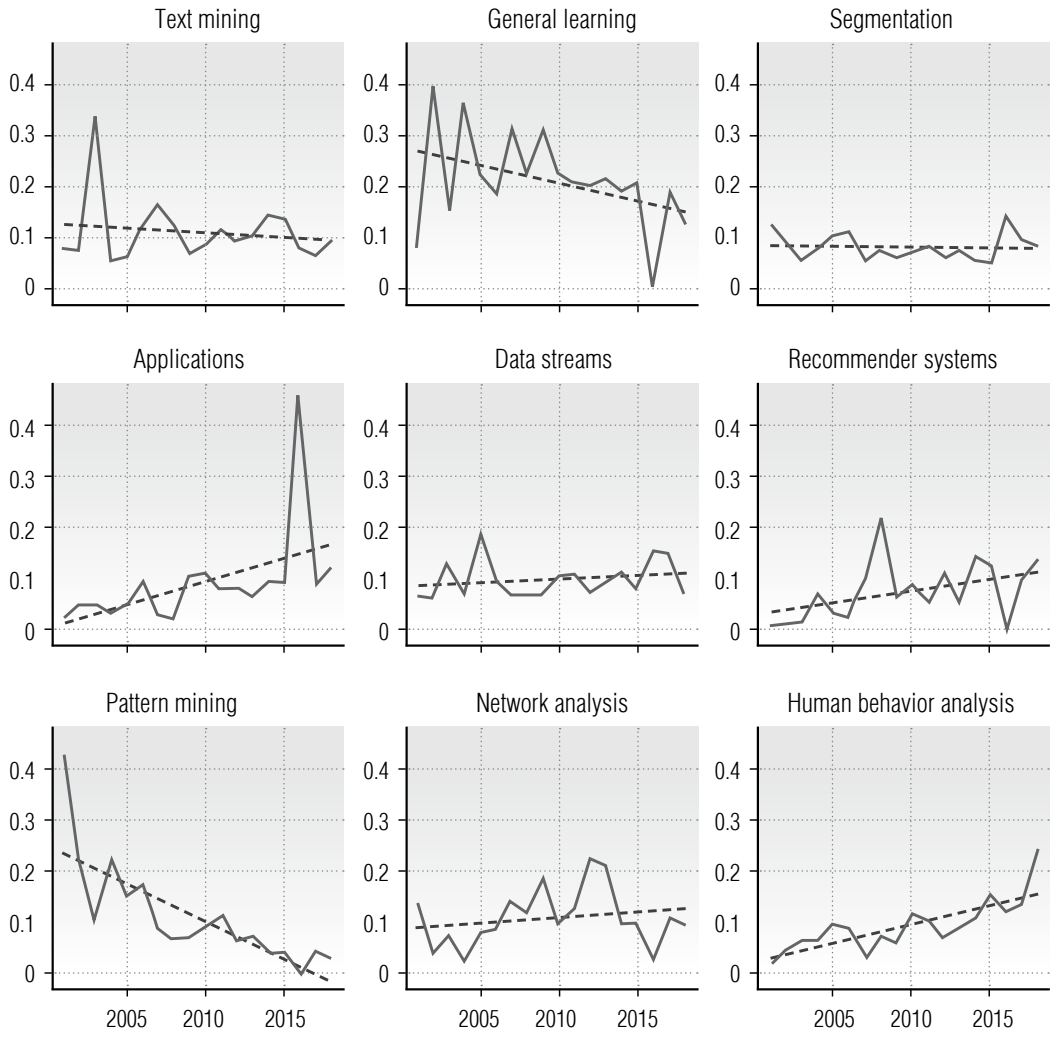

Fig. 5. The dynamics of the topics' popularity (solid line) and the trend (dashed line)

Behavior Analysis also has relatively strong connections with other directions. Recommender Systems, Network Analysis, and Text Mining are more isolated topics in our model because they are based on specialized algorithms.

The next stage of the analysis is a study of the dynamics of popularity and the impact of identified topics. Popularity is the derivative of the number of publications, and the impact is computed using the number of citations. *Figure 5* presents the dynamics of the topics' popularity (solid line), according to (2). The dashed line shows the trend. *Figure 6* presents the same data for the topics' impact, according to (3). These data shed light on the drift of interests of the data mining community regarding each research area.

Please note that the popularity and influence of many topics are subject to significant fluctuations. On the one hand, this can be explained by a short-term shift in the attention of researchers to hot topics. On the other hand, ICDM, although it is one of the most representative forums in the field of Data Mining, may not fully reflect the real dynamics of this discipline. For example, the word-

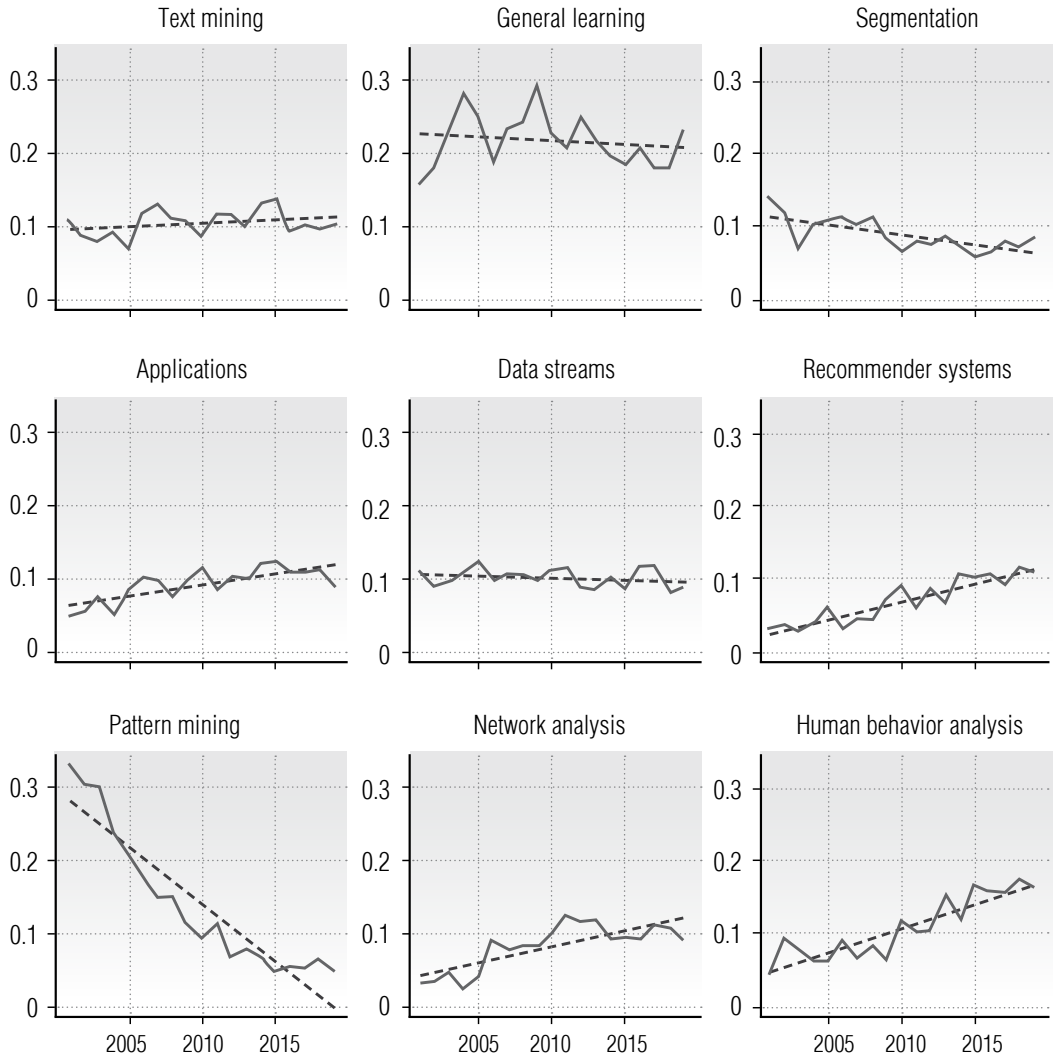

Fig. 6. The dynamics of the topics' impact (solid line) and the trend (dashed line)

ing of call for papers and the conference tracks by a program committee may affect researchers who submit works. However, we believe that an analysis of publications for 2001–2019 allows us to identify global trends, which is the goal of our study.

The data presented show that research attention to Pattern Mining is sharply decreasing both in terms of popularity and impact. The same, but less pronounced tendency is also characteristic of General Learning and Segmentation. These trends require more in-depth study. Firstly, it can be assumed that this is because the achievements in the field of machine learning algorithms and related techniques, including associations mining, are already so outstanding that further advancement requires serious efforts. Today, the most activity is observed in the field of deep learning, though, according to our results and analysis in [5, 6], deep neural networks are not directly related to Data Mining.

At the same time, the popularity of research related to Recommender Systems, Network Analysis, and Human Behavior Analysis is growing, which is most likely due to the increas-

ing availability of data and the practical value of these topics. The research direction related to practical Applications of data mining also is tending to grow. This can also explain the decline in interest in foundational algorithms; a significant part of the research community is focusing on more relevant practical issues.

The last two topics, Text Mining, and Data Streams have attracted steady interest from researchers. The results presented shed light on the structure and dynamics of data mining over the past twenty years and allow us to expand our understanding of this scientific discipline.

The next issue that is of interest from analyzing the content of publications is the diversity of documents. By analogy with (1), we can determine the diversity of a document through the entropy of its topics:

$$H_d = -\sum_i^T \theta_{di} \ln \theta_{di} ,$$

where θ_{di} is the weight of the topics i in document d;

T is the number of topics.

Figure 7 presents the mean diversity of proceedings of ICDM for 2001–2019. We see that the topic diversity of documents has grown steadily since the first conference and peaked in 2015. Over the past four years, there has been a reduction in the number of topics covered in one document.

We believe that this can be explained as follows. In the early 2000s, the main interest of researchers was focused on the knowledge discovery algorithms, which is presented by a list of critical topics highlighted in [5] and confirmed in [6] (*Table 1*). As they matured, these algorithms expanded their applications. Consequently, the set of topics covered in one scientific publication became more and more widespread. This can be considered as a search in the topic space that peaked in 2015. After 2015, a new research agenda was formed. As shown above, General Learning algorithms, as well as related areas such as Pattern Mining and Segmentation, are shifting to the background, although they continue to play an important role. More practical applications related to human behavior analysis, recommender systems, analysis of network communities, etc., come to the fore.

Table 3 presents a comparison of the data mining topics detected in our work and in [5] and [6]. Most of the topics of 2010 concentrate in the direction of General Learning. We

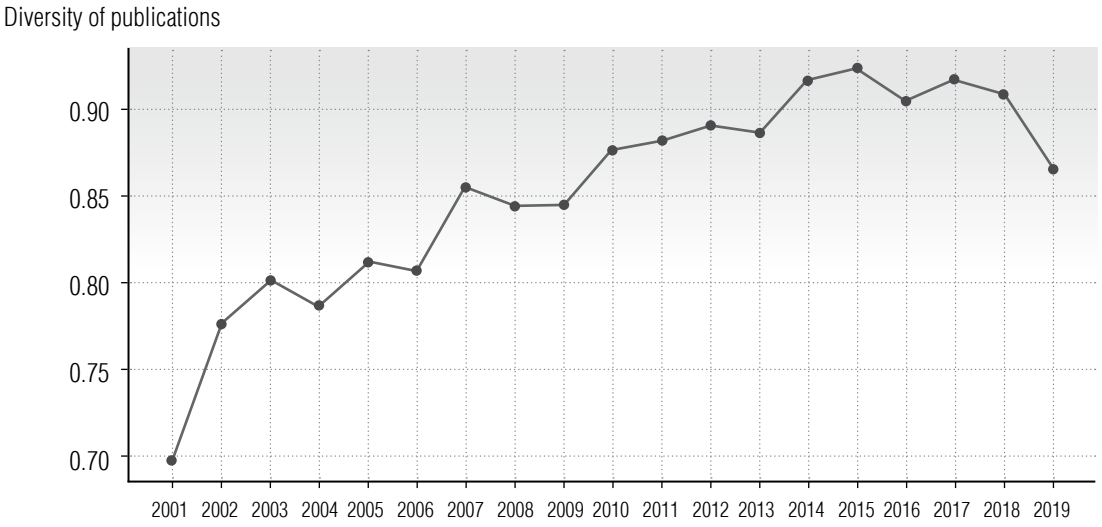

Fig. 7. The diversity of ICDM proceedings for 2001–2019

Table 3.

Mapping Data Mining topics

Data Mining Topics in 2020 (this work)	Data Mining Topics in 2010 [5, 6]
Text Mining	Link mining (e.g. PageRank algorithm)
General Learning	Classification (including C4.5, CART, kNN, and Naïve Bayes)
	Statistical learning (SVM and mixture models)
	Bagging and boosting
	Integrated mining (e.g. integrating classification and association rule mining)
	Rough sets
Segmentation	Clustering
Applications	NA
Data Streams	Sequential patterns
Recommender systems	NA
Pattern Mining	Association analysis
Network Analysis	Graph mining
Human Behavior Analysis	NA

included rough sets also in this category since this theory is applied to the classical knowledge discovery problems, such as discovering patterns in missing data [40].

New topics, the rapid growth of which our model identified, were not considered at all in 2010. We also note that the Text Mining topic detected in our work includes a much larger range of technologies and applications than searching for relations between documents.

Interestingly, in the middle of 2010s, researchers in the field of economics and management recorded an increase in interest in Data-Driven Decision Making (DDD) [41]. That refers to the practice of basing decisions on the analysis of data rather than purely on human knowledge and intuition. Authors of [42] report that the use of DDD in US manufacturing nearly tripled (from 11 percent to 30 percent of plants) between 2005 and 2010. More recent studies confirm the increasing role of DDD as one of the best management practices [43].

The description of DDD which is given in management literature entirely coincides with the definition of Data Mining discussed at the beginning of our work. DDD also includes the engineering and processing of data and the discovery of useful patterns. However, DM considers this activity from a technological point of view. DDD approaches this issue in the context of closely related processes in the organization, including purely human activities. Nevertheless, we can consider the growing interest in DDD as one of the key drivers affecting the shift in DM studies that is presented in *Figure 7*.

Conclusion

We presented a study of the intellectual structure of Data Mining as a scientific discipline carried out using topic analysis. This approach made it possible to identify nine main areas in Data Mining and to study their dynamics.

The main result of our work is that we have discovered a shift in interests from machine learning algorithms to more practical applications. According to our data, this change of focus took shape in the middle of the 2010s. We attribute this shift to a combination of three factors:

Firstly, the basic data mining algorithms have reached a high level of maturity.

Secondly, with the development of social networks, a large amount of data has become available.

Third, there has been a steady demand from the business for data-driven decision-making. ■

References

1. Piatetsky-Shapiro G., Fayyad U. (2012) An introduction to SIGKDD and a reflection on the term 'data mining'. *ACM SIGKDD Explorations Newsletter*, vol. 13, no 2, pp. 102–103. DOI: 10.1145/2207243.2207269.
2. Witten I.H., Frank E., Hall M., Pal C. (2017) *Data mining: Practical machine learning tools and techniques.* Cambridge, MA: Morgan Kaufmann.
3. Han J., Kamber M., Pei J. (2012) *Data mining: Concepts and techniques.* Waltham, MA: Morgan Kaufmann. DOI: 10.1016/C2009-0-61819-5.
4. Rather B. (2011) *Statistical and machine-learning data mining: Techniques for better predictive modeling and analysis of big data*. Sound Parkway, NW: CRC Press.
5. Yang Q., Wu X. (2006) 10 challenging problems in data mining research. *International Journal of Information Technology & Decision Making*, vol. 5, no 4, pp. 597–604. DOI: 10.1142/S0219622006002258.
6. Wu X. (2010) 10 years of data mining research: retrospect and prospect. Proceedings of the *10th IEEE International Conference on Data Mining (ICDM), Sydney, Australia, 13–17 December 2010*, p. 7. DOI: 10.1109/ICDM.2010.172.
7. Liao S.H., Chu P.H., Hsiao P.Y. (2012) Data mining techniques and applications – A decade review from 2000 to 2011. *Expert Systems with Applications*, vol. 39, no 12, pp. 11303–11311. DOI: 10.1016/j.eswa.2012.02.063.
8. Van Eck N.J., Waltman L. (2009) Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, vol. 84, no 2, pp. 523–538. DOI: 10.1007/s11192-009-0146-3.
9. Thilakaratne M., Falkner K., Atapattu T. (2019) A systematic review on literature-based discovery: general overview, methodology, & statistical analysis. *ACM Computing Surveys*, vol. 52, no 6, article no 129. DOI: 10.1145/3365756.
10. Zelenkov Y. (2019) The topic dynamics in knowledge management research. Proceedings of the *14th International Conference on Knowledge Management in Organizations (KMO 2019), Zamora, Spain, 15–18 July 2019*, pp. 324–335. DOI: 10.1007/978-3-030-21451-7_28.
11. Mann G.S., Mimno D., McCallum A. (2006) Bibliometric impact measures leveraging topic analysis. Proceedings of the *6th ACM/IEE Joint Conference on Digital Libraries (JCDL '06), Chapel Hill, NC, USA, 11–15 June 2006*, pp. 65–74. DOI: 10.1145/1141753.1141765.
12. Dam H.K., Ghose A. (2016) Analyzing topics and trends in the PRIMA literature. Proceedings of the *19th International Conference on Principles and Practice of Multi-Agent Systems (PRIMA* 2016*), Phuket, Thailand, 22–26 August 2016*, pp. 216–229. DOI: 10.1007/978-3-319-44832-9_13.
13. Dias L., Gerlach M., Scharloth J., Altman E.G. (2018) Using text analysis to quantify the similarity and evolution of scientific disciplines. *Royal Society Open Science*, vol. 5, no 1, article no 171545. DOI: 10.1098/rsos.171545.
14. Jensen S., Liu X., Yu Y., Milojevic S. (2016) Generation of topic evolution trees from heterogeneous bibliographic networks. *Journal of Informetrics*, vol. 10, no 2, pp. 606–621. DOI: 10.1016/j.joi.2016.04.002.
15. Syed S., Spruit M. (2017) Full-text or abstract? Examining topic coherence scores using latent Dirichlet allocation. Proceedings of the *4th IEEE International Conference on Data Science and Advanced Analytics (DSAA 2017), Tokyo, Japan, 19–21 October 2017*, pp. 165–174. DOI: 10.1109/DSAA.2017.61.
16. Blei D.M., Ng A.Y., Jordan M.I. (2003) Latent Dirichlet allocation. *Journal of Machine Learning Research*, no 3, pp. 993–1022.
17. Mockus J. (2012) *Bayesian approach to global optimization: Theory and applications.* Heidelberg: Springer. DOI: 10.1007/978-94-009-0909-0.
18. Tang J., Meng Z., Nguyen X., Mei Q., Zhang M. (2014) Understanding the limiting factors of topic modeling via posterior contraction analysis. *Proceedings of Machine Learning Research*, vol. 32, no 1, pp. 190–198.

19. Molina L.C., Belanche L., Nebot A. (2002) Feature selection algorithms – A survey and experimental evaluation. Proceedings of the *2nd IEEE International Conference on Data Mining (ICDM* 2002*), Maebashi City, Japan, 9–12 December 2002*, pp. 306–313. DOI: 10.1109/ICDM.2002.1183917.

20. Read J., Pfahringer B., Holmes G. (2008) Multi-label classification using ensembles of pruned sets. Proceedings of the *8th IEEE International Conference on Data Mining (ICDM), Pisa, Italy, 15–19 December 2008*, pp. 995–1000. DOI: 10.1109/ICDM.2008.74.

21. Chen X., Pan W., Kwok J.T., Carbonell J.G. (2009) Accelerated gradient method for multi-task sparse learning problem. Proceedings of the *9th IEEE International Conference on Data Mining (ICDM), Miami Beach, FL, USA, 6–9 December 2009*, pp. 746–751. DOI: 10.1109/ICDM.2009.128.

22. Shin K. (2011) Partitionable kernels for mapping kernels. Proceedings of the *11th IEEE International Conference on Data Mining (ICDM), Vancouver, BC, Canada, 11–14 December 2011*, pp. 645–654. DOI: 10.1109/ICDM.2011.115.

23. Li Z., Yada K. (2015) Why do retailers end price promotions – A study on duration and profit effects of promotion. Proceedings of the *15th IEEE International Conference on Data Mining Workshop (ICDMW), Atlantic City, NJ, USA, 14–17 November 2015*, pp. 328–335. DOI: 10.1109/ICDMW.2015.56.

24. Camino R.D., State R., Montero L., Valtchev P. (2017) Finding suspicious activities in financial transactions and distributed ledgers. Proceedings of the *17th IEEE International Conference on Data Mining Workshops (ICDMW), New Orleans, LA, USA, 18–21 November 2017*, pp. 787–796. DOI: 10.1109/ICDMW.2017.109.

25. Guo T., Bifet A., Antulov-Fantulin N. (2018) Bitcoin volatility forecasting with a glimpse into buy and sell orders. Proceedings of the *18th IEEE International Conference on Data Mining (ICDM), Singapore, 17–20 November 2018*, pp. 989–994. DOI: 10.1109/ICDM.2018.00123.

26. Gouda K., Zaki M.J. (2001) Efficiently mining maximal frequent itemsets. Proceedings of the *2001 IEEE International Conference on Data Mining (ICDM), San Jose, CA, USA, 29 November – 2 December 2001*, pp. 163–170. DOI: 10.1109/ICDM.2001.989514.

27. Ma S., Hellerstein J.L. (2001) Mining mutually dependent patterns. Proceedings of the *2001 IEEE International Conference on Data Mining (ICDM), San Jose, CA, USA, 29 November – 2 December 2001*, pp. 409–416. DOI: 10.1109/ICDM.2001.989546.

28. Zhou D., Orshanskiy S.A., Zha H., Gees C.L. (2007) Co-ranking authors and documents in a heterogeneous network. Proceedings of the *7th IEEE International Conference on Data Mining (ICDM), Omaha, NE, USA, 28–31 October 2007*, pp. 739-744. DOI: 10.1109/ICDM.2007.57.

29. Tang J., Jin R., Zang J. (2008) A topic modeling approach and its integration into the random walk framework for academic search. Proceedings of the *8th IEEE International Conference on Data Mining (ICDM), Pisa, Italy, 15–19 December 2008*, pp. 1055–1060. DOI: 10.1109/ICDM.2008.71.

30. Shehata S., Karray F., Kamel M. (2006) Enhancing text clustering using concept-based mining model. Proceedings of the *6th IEEE International Conference on Data Mining (ICDM), Hong Kong, China, 18–22 December 2006*, pp. 1043–1048. DOI: 10.1109/ICDM.2006.64.

31. Yang D., Li B., Rettig L., Cudre-Mauroux P. (2017) HistoSketch: Fast similarity-preserving sketching of streaming histograms with concept drift. Proceedings of the *17th IEEE International Conference on Data Mining (ICDM), New Orleans, LA, USA, 18–21 November 2017*, pp. 545–554. DOI: 10.1109/ICDM.2017.64.

32. Dhurandhar A. (2010) Learning maximum lag for grouped graphical Granger models. Proceedings of the *10th IEEE International Conference on Data Mining Workshops (ICDMW), Sydney, Australia, 13 December 2010*, pp. 217–224. DOI: 10.1109/ICDMW.2010.9.

33. Wang H., Li Z., Lee W.-C. (2014) PGT: Measuring mobility relationship using personal, global and temporal factors. Proceedings of the *14th IEEE International Conference on Data Mining (ICDM), Shenzhen, China, 14–17 December 2014*, pp. 570–579. DOI: 10.1109/ICDM.2014.111.

34. Pirbhulal S., Wu W., Li G. (2018) A biometric security model for wearable healthcare. Proceedings of the *18th IEEE International Conference on Data Mining Workshops (ICDMW), Singapore, 1 7–20 November 2018*, pp. 136–143. DOI: 10.1109/ICDMW.2018.00026.

35. Yang J., Leskovec J. (2012) Defining and evaluating network communities based on ground-truth. Proceedings of the *12th IEEE International Conference on Data Mining (ICDM), Brussels, Belgium, 10–13 December 2012*, pp. 745–754. DOI: 10.1109/ICDM.2012.138.

36. Shi L., Tong H., Tang J., Lin C. (2014) Flow-based influence graph visual summarization. Proceedings of the *14th IEEE International Conference on Data Mining (ICDM), Shenzhen, China, 14–17 December 2014*, pp. 983–988. DOI: 10.1109/ICDM.2014.128.

37. Hung M.-C., Yang D.-L. (2001) An efficient fuzzy c-means clustering algorithm. Proceedings of the *2001 IEEE International Conference on Data Mining (ICDM), San Jose, CA, USA, 29 November – 2 December 2001*, pp. 225–232. DOI: 10.1109/ICDM.2001.989523.

38. Pei Y., Zaiane O.R., Gao Y. (2006) An efficient reference-based approach to outlier detection in large datasets. Proceedings of the *6th IEEE International Conference on Data Mining (ICDM), Hong Kong, China, 18–22 December 2006*, pp. 478–487. DOI: 10.1109/ICDM.2006.17.

39. Wang X., Chen Y., Yang J., Wu L., Wu Z., Xie X. (2018) A reinforcement learning framework for explainable recommendation. Proceedings of the *18th IEEE International Conference on Data Mining (ICDM), Singapore, 17–20 November 2018*, pp. 587–596. DOI: 10.1109/ICDM.2018.00074.

40. Wang H., Wang S. (2009) Discovering patterns of missing data in survey databases: an application of rough sets. *Expert Systems with Applications*, vol. 36, no 3, part 2, pp. 6256–6260. DOI: 10.1016/j.eswa.2008.07.010.

41. Provost F., Fawcett T. (2013) Data science and its relationship to big data and data-driven decision making. *Big Data*, vol. 1, no 1, pp. 51–59. DOI: 10.1089/big.2013.1508.

42. Brynjolfsson E., McElheran K. (2016) The rapid adoption of data-driven decision-making. *American Economic Review*, vol. 106, no 5, pp. 133–139. DOI: 10.1257/aer.p20161016.

43. Song P., Zheng C., Zhang C., Yu X. (2018). Data analytics and firm performance: An empirical study in an online B2C platform. *Information & Management*, vol. 55, no 5, pp. 633–642. DOI: 10.1016/j.im.2018.01.004.

About the authors

Yury A. Zelenkov

Dr. Sci. (Tech.);

Professor, Department of Business Informatics, Graduate School of Business, National Research University Higher School of Economics, 20, Myasnitskaya Street, Moscow 101000, Russia;

E-mail: yzelenkov@hse.ru

ORCID: 0000-0002-2248-1023

Ekaterina A. Anisichkina

Student, BSc Program "Business Informatics", Graduate School of Business, National Research University Higher School of Economics, 20, Myasnitskaya Street, Moscow 101000, Russia;

E-mail: eaanisichkina@edu.hse.ru