# Dynamics of Data Mining Research Trends: A Two-Decade Review Using Topic Analysis

Yuri Zelenkov
Graduate School of Business
National Research University Higher School of Economics
Moscow, Russia
yuri.zelenkov@gmail.com

Ekaterina Anisichkina
Graduate School of Business
National Research University Higher School of Economics
Moscow, Russia
catherine.aniss@gmail.com

***Abstract*— The work analyses the intellectual structure of data mining as a scientific discipline. To do this, we use topic analysis (namely, Latent Dirichlet Allocation) 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 declines last time. The analysis also showed that the attention to topics such as *Pattern Mining* (detecting associations) and *Segmentation* (object separation algorithms such as clustering) decreases. 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 increasing availability of data and the practical value of these topics. The research direction related to practical *Applications* of data mining also tends to grow. The last two topics, *Text Mining* and *Data Streams* have attracted steady interest from researchers. The presented results 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. 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.**

***Keywords—data mining topics, topic analysis, scientometrics.***

## I. Introduction

The term "data mining" 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 (DM) 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 includes 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. The dynamic of these topics (for example, a change in the number of publications and their citation) reflects a shift in the interests of the research community. In particular, the study of this dynamic allows determining the most relevant areas of research in the present and extrapolate them in the near future. In addition, the understanding of fundamental shifts in the interests of researchers helps 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 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.

## II. Data

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 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 5121 publications of the main ICDM tracks and related workshops.

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

## III. Research Method

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

Let there is 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

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

So, 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 5121 papers downloaded from the Web of Science database, as described in the previous section. According to the [5], 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 with dimensions of 5121 x 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 additional assumption that the distribution $\Theta$ of documents $\theta_d$ and distribution $\Phi$ of topics $\varphi_t$ are spawn by Dirichlet distributions. To build the model, one should define a number of topics $|T|$; the LDA algorithm computes distributions $\Theta$ and $\Phi$. As a result, each topic is presented by the weighted list of words; the weight of 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. We used more advanced techniques, namely, Bayesian optimization. Such an approach allows us to optimize simultaneously not only the number of the topics and also parameters of $\Theta$ and $\Phi$ distributions and other parameters of the algorithm. Optimization target is a perplexity that is be computed as

$$P(D) = \exp\left[\frac{1}{2}\sum_{d \in D}\sum_{w \in d} n_{dw} \ln p(w|d)\right].$$

The perplexity of collection $D$ is a measure of the language quality and 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.

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

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

here $n_w$ 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), its 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 $\theta_{dt}$ 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

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

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

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

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

## IV. Results and Discussion

Performing all preprocessing operation 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 (Fig. 1), 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).

So, 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. Table 2 lists the topics description.

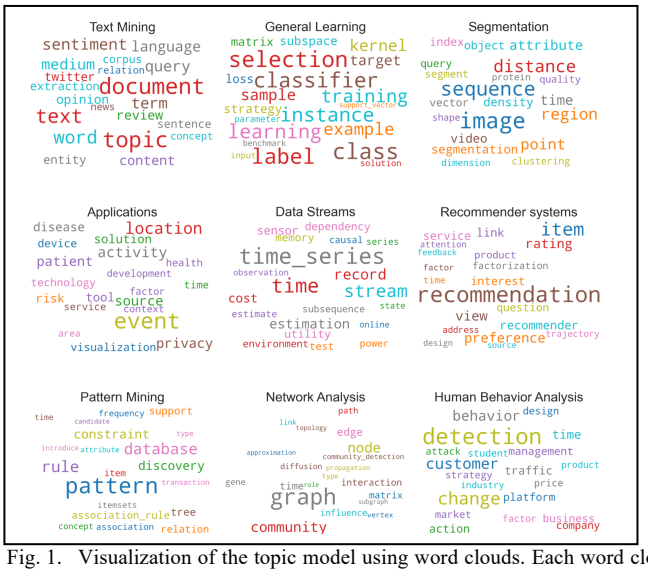


Fig. 1. Visualization of the topic model using word clouds. Each word cloud represents one detected topic where the size of words indicates the relevance of each word to that particular topic.

TABLE I. 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 |

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.

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.

$$H_d = -\sum_{i}^{t} \theta_{di} \ln \theta_{di}.$$

Here $\theta_{di}$ is the weight of the topics $i$ in document $d$; $t$ is the number of topics.

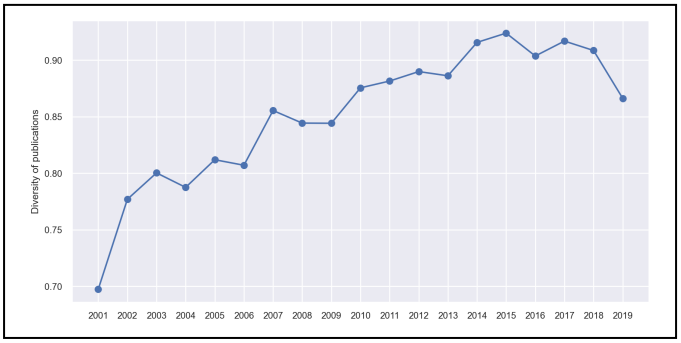


Fig. 2. The diversity of ICDM Proceedings for 2001-2019.

Fig. 2 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 2014. Over the past five 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. 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 2014. After 2014, 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.

## REFERENCES


[1] G. Piatetsky-Shapiro, and U. Fayyad, "An introduction to SIGKDD and a reflection on the term 'data mining'," ACM SIGKDD Explorations Newsletter, vol. 13, no. 2, pp. 102-103, 2012.

[2] I. H. Witten, E. Frank, M. Hall, and C. Pal, Data Mining: Practical Machine Learning Tools and Techniques, 4rd ed. Cambridge, MA: Morgan Kaufmann, 2017.

[3] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed. Waltham, MA: Morgan Kaufmann, 2012.

[4] B. Rather, Statisitcal and Machine-Learning Data Mining: Techniques for Better Predictive Modeling and Analysis of Big Data, 2nd ed. Sound Parkway, NW: CRC Press, 2011.

[5] S. Syed, and M. Spruit, "Full-text or abstract? Examining topic coherence scores using latent Dirichlet allocation," in IEEE Intl. Conf. on Data Science and Advanced Analytics (DSAA), 2017 pp. 165-174.