

DOI 10.36074/grail-of-science.29.04.2022.006

KNOWLEDGE DISCOVERY AND DATA MINING OF STRUCTURED AND UNSTRUCTURED BUSINESS DATA: PROBLEMS AND PROSPECTS OF IMPLEMENTATION AND ADAPTATION IN CRISIS CONDITIONS

Maxim Krasnyuk 

Ph.D., Associate Professor

Department of Finance

The National University of Food Technologies, Ukraine

Yurii Kulynych 

Ph.D. of Economics, Associate Professor

Department of Finance

Institute of Economics and Management National University of Food Technologies, Ukraine

Svitlana Krasniuk 

Senior Lecturer

Institute of Law and Modern Learning Technologies

Kyiv National University of Technologies and Design, Ukraine

Summary. *In modern conditions of the development of the global economy and in connection with the emergence of new branches of economic activity in the field of IT, the phenomenon of Structured and Unstructured Big Data - the use of Data Science for advanced in-depth analysis of data and knowledge in all possible modes - leads to competitive advantages for corporations and institutions, both at the regional and interstate levels, which is especially relevant in the context of the current macroeconomic and military crisis [1].*

The following topical issues are systematically investigated in the article: current status and prospects for further development of Knowledge Discovery in Data Base (KDD), problems and critical issues of theory and practice of Data Mining, the specifics of effective use of Knowledge Discovery in DB (Data Base) in the current crisis in Ukraine.

The above trends and features of the KDD market should be taken into account in further theoretical research and practical implementation or reengineering of KDD systems in Ukraine. The obtained results are relevant and applicable not only for local companies and organizations, but also for international applications in the context of global, regional macroeconomic and current national crisis phenomena.

Keywords: *Data Mining, Knowledge Discovery, Structured and Unstructured Data, Big Data, Crisis Management.*

Introduction.

Most enterprises and organizations (commercial, industrial, scientific, service, etc.) have been registering and recording huge amounts of heterogeneous information (quantitative, qualitative, textual, multimedia, etc.) on all aspects of their activities for years.

However, it is useless to hope that these data sets will significantly help companies to make quality strategic decisions without appropriate technologies, algorithms and scenarios for data preparation, processing and analysis. After all, in order to formulate the correct premises for making effective management decisions, we need objective knowledge about the nature, relationships and patterns of the studied subject area. Finding the necessary facts in a database or repository is not so difficult, but in today's information society not only the facts themselves are needed, but especially new knowledge.

The importance of knowledge for the organization and their further effective use has been recognized for many years by leading theorists and practitioners of management, but the very concept of knowledge management requires a rethinking of traditional managerial thought [2].

One of the definitions of the term knowledge management is to improve the effectiveness of the company by improving the structure, discipline and practical activities for the collection and processing of knowledge in the corporation and providing them for collective use. Intelligent technologies and systems for Knowledge Discovery and Data Mining are an integral part of this knowledge-oriented management concept.

Problem statement and relevance of the research.

In the information economy, effective, efficient analytical data processing is a very important issue in order to obtain potentially useful and understandable information and knowledge to make an effective decision. Due to the following modern specifics: the data are unlimited; data are heterogeneous (quantitative, qualitative, textual); the results of the analysis must be specific and clear; tools for such data processing should be easy to use.

The active and innovated development and spreading of IT has led to an avalanche of generated, accessible and stored structured and semi-structured data (from corporate Data Bases and Data Warehouses, OLTP data, Web data, Social networks data, IoT, etc.) and the corresponding development of updated algorithms and technologies of Big Data analytics. Moreover, these technologies, tools and means must be effective in the conditions of continuous data growth and their distribution across computer network nodes [3].

However, in emerging markets, only a small proportion of stored data (including hypertext data, streaming multimedia data, hybrid and metadata) is deeply analyzed by modern technologies and methods.

Given the above prerequisites, both government agencies and commercial enterprises - the question of accelerated and effective implementation and / or adaptation of KDD services arose.

Knowledge Discovery in DB or Data Mining is the detection of structured and unstructured data, previously unknown or hidden, available for further interpretation, practical useful patterns and patterns (knowledge) in primary,

accumulated as a result of business transactions processing in order to form an appropriate knowledge base and acceptance sound and optimal management and / or business decisions.

KDD systems use sophisticated analysis and modeling to find models and relationships hidden in the Data Warehouse — models that cannot be effectively found by classical statistical analysis methods.

Thus, KDD is a multidisciplinary industry that combines extensive mathematical tools and the latest advances in information technology. KDD technology harmoniously combines formalized methods and methods of informal analysis, methods for taking uncertainty into account in data and knowledge.

This determines the relevance of the scientific field of KDD, which involves intellectual analysis of accumulated data in an automated mode to identify new, hidden patterns, and forming a corporate knowledge base from them, taking into account the current crisis phenomena [4].

The general theoretical issues of KDD algorithms and methods are widely studied in the modern Western scientific literature, but the following remain insufficiently resolved and relevant issues: problems and critical issues of effective implementation and use of KDD systems; strategic and tactical perspectives and features of further development and transformation of KDD; regional and industry specifics of the KDD market.

The above problematic issues prove the relevance of this work, as they must be taken into account in the future implementation and / or adaptation of KDD systems in Ukraine in modern national and global crisis conditions.

The modern global economy is characterized not only by processes such as the globalization of markets, the internationalization of enterprises, the rapid development of innovative information technologies, but also the rapid spread of crisis phenomena between cooperative economies, increasing global demands for total business efficiency, and they all put forward new demands for corporate management [5].

It can be supposed, that in today's global economy conditions (in connection with the emergence of new sectors of economic activity), application of technologies and algorithms, optimal and effective scenarios of their implementation for Knowledge Discovery and Data Mining, - leads not only to additional competitive advantages by companies and organizations (and hence increase their investment attractiveness and capitalization), but also serves as an important anti-crisis factor in emerging markets in modern conditions [6].

The main part and results.

However, it is first necessary to present the results of the study of general problems of KDD technology and critical issues related to the practical application in developed markets:

- The technology of finding new knowledge in the data bases can not answer the questions that were not asked. It does not replace analysts or managers, but gives them a modern, powerful tool to improve the work they do.

- High percentage of erroneous results. KDD helps analysts find models and relationships in data, but it does not talk about the value of these models to the organization. Each model must be tested in a real environment. KDD, unfortunately,

very often generates many misleading and insignificant discoveries. Many users and analysts claim that KDD tools can produce thousands of erroneous, statistically inaccurate or meaningless results. In this case, the user must understand which of the results make real sense. Some scholars warn that conventional KDD methods only simplify the absurdly complex art of analysis and may lead to incorrect conclusions. Practitioners say that strange patterns are often found; in 99.9% of cases they are wrong, because they can be based on only a few random examples. And KDD packages usually do not have built-in verification tools.

- The complexity of the tools. KDD requires the user to understand the operation of the tools used and the algorithms on which it is based. Complexity is a significant barrier to the implementation of KDD. In fact, KDD is the result of the joint efforts of experts in three areas. Project management should be undertaken by business professionals whose task is to form a set of business goals (business objectives) and the subsequent interpretation of the results. An developer-analyst who understands KDD methods, statistics and tools must create a reliable model. And information technology specialists provide data processing as well as technical support.

- User experience. Different KDD tools have their strengths and weaknesses. Therefore, specific programs must clearly correspond to the level of preparedness of the user and his specific goals. In addition, KDD usually involves the use of some technical jargon, which can make it very difficult for the inexperienced user to understand the program, its essence, practical results, as well as which product and which method is best used to achieve certain business goals. Therefore, often a potential customer may refuse to use KDD altogether. Even worse, if a client invests a lot of money and goes the wrong way or spends money on various tools in order to finally understand how to apply KDD in this area. The use of KDD must be inextricably linked to user development. However, there are too few KDD specialists who are well-versed in a particular field of economics. As stated above, extracting useful data or knowledge is impossible without understanding the nature of the data, and in many cases a careful interpretation of the dependencies or patterns that have been identified is required. Therefore, working with KDD tools requires close collaboration between a business expert and a KDD tool specialist.

- Inconsistency of forecasting results with the real situation. There is one challenge facing KDD that many experts believe is unsolvable and that justifies the skepticism that is often heard in this niche market. KDD tools predict consumer behavior based on data from past periods, i.e. provide information that a person, based on his previous purchases, demographics and other parameters, will want to buy with the highest probability. But, critics say, KDD will never clearly predict what a person will actually want to buy. And to increase profitability (i.e. achieve the main goal of marketing), first of all, we need not so much to find out what a person is satisfied with now, how much to find out, and what he/she will buy with the greatest desire. The only way to solve this problem is to ask customers what they really want, rather than relying on information about the nature of their past purchases.

- Large labor costs. The results of KDD largely depend on the level of data preparation, rather than on the capabilities of a particular algorithm. About 75% of the labor costs of the KDD problem are collected, cleaned, properly prepared, the

problem set (before the algorithms are executed) and further testing and adjustment of forecasting models.

- Confidentiality. The KDD analytical process itself applies to accumulated anonymous data, revealing usability, acquisition trends, and dozens, if not hundreds, of other factors. However, the next step in data processing - trying to relate it to the nature of a particular client's behavior in order to gain some personal experience of interacting with that person - is a cause for concern among privacy advocates. Privacy advocates talk about how important it is to be honest with customers and tell them about the data collected and the purposes for which it is used. To do this, directives have been developed that seem simple, but in practice, their implementation is difficult. In general, they sound like this: tell the people you collect data about and how you plan to use it; give the opportunity not to specify this information about yourself; provide review and correction of personal records. KDD software vendors have decided to take the hard way of ensuring privacy by promoting the ability of customers to comply with these guidelines.

- Efficiency and productivity. There are several science schools, each with its own views on KDD technology and its effectiveness. Some marketers and application vendors consider KDD tools, which rarely work online, to be an outdated technology. In this form, KDD can be used to create a broad profile of certain types of customers, but these tools do not provide key information about the nature of a particular person's behavior. Other developers believe that KDD tools are too slow and cannot perform accurate analysis and offer the user the desired service while the user is on the provider's website. Particular doubts about the effectiveness arise against systems based on rules that perform KDD analysis on the server. The literature provides an example of one company that estimated that in order to reflect all the possibilities on their site, you will need to write 90 thousand rules for use in traditional KDD methods. The company decided to focus on writing one thousand rules in the first stage, given the high probability of error. However, it should be borne in mind that the rules are written by people, and they can be biased.

- Use of a special database. Of course, KDD vendors require the use of an expensive specialized database, storefronts, or data warehousing on a dedicated server that can efficiently research data.

- High cost. The professional KDD program costs between \$ 500,000 and \$ 1.5 million, which is needed for software, hardware and technical support. When investing in such a project, it is necessary to make sure that the efficiency of investment will be quite high. A good test is a small KDD project (from \$100,000 to \$200,000 dollars), which will determine whether the amount and quality of data available to make KDD useful for the company. The KDD market is growing. However, software tools make up only 15% of the budget of the whole project. Most of the money goes to service companies and system integrators, which "protect" users from the complexities of technology. This need for outside help and powerful equipment leads to an increase in the average cost of implementing KDD to \$ 2 million and more for a corporate project. With this money, suppliers often promote the following idea: KDD provides in-depth knowledge that constantly leads to "breakthroughs" in business. However, this is not always true.

As general perspectives of the KDD direction it is possible to note that:

- the market of KDD systems is developing exponentially. However, there is a real lag behind the capabilities of modern commercial data mining systems from theoretical advances in this area;

- KDD systems are independently developed in two main areas: a) as a mass product for business applications (increasing share in this market is gaining products of large companies - database providers, including Oracle and IBM); b) as tools for conducting unique research (genetics, chemistry, medicine, etc.);

- most KDD systems are designed to work with a single data source. In those rare cases where the system can work with several data sources at the same time, the basic source is still one; however, the system does not actually support the use of metadata from these sources;

- in the vast majority of systems studied by the author there is no possibility of distributed machine learning and learning from different local sources. This leads to the fact that learning tasks are weighed strictly consistently, which places very high demands on the hardware;

- despite the significant number of KDD methods, the priority is gradually shifting towards logical search algorithms in the data of if-then rules. Because the results of such algorithms are efficient and easy to interpret. It should be noted that the main problem of logical methods of identifying patterns is the problem of searching for options in a reasonable time. Known methods either artificially limit such a search (algorithms "KOPA", APRIORI etc.), or build decision trees (algorithms CART, CHAID, ID3, C5, Sipina, etc.), which have fundamental limitations on the effectiveness of finding rules;

- in universal KDD systems there are practically no mechanisms for structuring the found knowledge in the form of knowledge bases and tools to support decision-making based on the found knowledge;

- actual technology is still Web mining, which is a scalable modular information system for collecting information and managing content for a web-site. Web mining focuses on three main tasks: collecting the maximum amount of information about site visitors and requested resources; research of collected data; and generation of personalized, based on research results, content. All this should maximize the number of visitors to the website;

- the current trend is the development of Text Mining, which involves the integration of KDD tools (numerical information analysis tools) with methods of text analysis in natural language - Text Mining algorithms.

The use of KDD in emerging markets in crisis conditions is currently a topical issue. The following factors should be taken into account when analyzing such prospects and peculiarities of KDD application in the domestic economy:

- 1) relatively insignificant time of functioning of the researched sphere, and consequently, insufficient period of accumulation of input data in a data warehouse. The danger here is not so much in the inability to identify the desired relationships in the few data and to build models based on them, as in obtaining statistically insignificant models and making incorrect decisions based on them. Thus, constant monitoring of the statistical significance of KDD results is necessary;

- 2) widespread use of subjective factors (shadow economy) in the process of economic decision-making and insufficient implementation of Western corporate

standards. And according to KDD theory, built models should be interpretable and objective and transparent;

3) both constant changes in the official "rules of the game" at the macro level of Ukraine's economy, and the sudden emergence of new factors at the microeconomic level of a particular enterprise;

4) the existence in the market of interconnected systems, the interests of which may be conflicting;

5) the presence of non-systemic time lag (delay) in the transmission of information both within the economic system and when influencing it;

6) a certain unwillingness of the top management of domestic corporations to spend significant funds not only on analytical software, and even more so on KDD;

7) insufficient level of knowledge of managers and their psychological non-perception of statistics and artificial intelligence, so they do not sufficiently support and cannot directly use data mining systems that require complex debugging or special data preparation. Because KDD tools cannot work without user support: those who have a good understanding of the business field, the data itself and the general nature of the analytical methods used.

The above factors may give reason to doubt the correctness and significance of the knowledge gained from the KDD system.

The above trends and features of the KDD market should be taken into account in further theoretical research and practical implementation of KDD systems in Ukraine.

Conclusions and perspectives of further research.

When considering the possible use of KDD, it is important for the management of Ukrainian companies to realize that the means of intelligent computing are a real way to increase efficiency. However, it should be borne in mind that building a model is only one step in the process of finding new knowledge. To obtain correct results, it is necessary to collect and prepare data and test the model in the real world. The best model can be found after building models of different types on different technologies [10].

Thus, the question is not whether new technologies are needed, but how to apply them in each case. The cost of setting the task and maintaining intelligent systems can be an order of magnitude higher than the cost of a single software package. Obviously, it is worth spending part of the money on training specialists - in the end it will be cheaper and more efficient. The role of specialized consulting firms that provide comprehensive project support, including problem diagnostics, analysis of solution methods, development of recommendations, implementation of the chosen approach, support, optimization is growing.

References:

- [1] Hrashchenko I. and Krasniuk S. (2015) Problems of regional development of Ukraine under globalisation process. *Visnyk Mizhnarodnoho humanitarnoho universytetu*. Serii: Ekonomika i menedzhment, 2015. - №11. – p. 26-32
- [2] Krasnyuk M.T. (2006) Problemy zastosuvannya system upravlinnia korporatyvnymy znanniamy ta yikh taksonomiia [Problems of applying corporate knowledge management systems and their taxonomy] Modeliuvannya ta informatsiini systemy v ekonomitsi: Mizhvid. nauk. zb. Zasnov. U 1965 r. Vyp. 73 / Vidp. red. V.K. Halitsyn. - K.:KNEU, 2006. – 256 s. [in Ukrainian]

- [3] Krasnyuk M.T., Hrashchenko I.S., Kustarovskiy O.D. and Krasniuk S.O. (2018) Methodology of effective application of Big Data and Data Mining technologies as an important anti-crisis component of the complex policy of logistic business optimization // *Economies' Horizons*, No. 3(6), pp. 121–136, doi: [https://doi.org/10.31499/2616-5236.3\(6\).2018.156317](https://doi.org/10.31499/2616-5236.3(6).2018.156317)
- [4] Krasnyuk M.T. and Kustarovskiy O.D. (2017) Problemy ta perspektyvy rozvytku ukraïnskykh lohistychno-informatsiinykh system v umovakh hlobalizovanoi ekonomiky ta makroekonomichnykh kryzovykh yavlyshch [The problems are the prospect of developing Ukrainian logistical and informational systems in the global economy and the macroeconomic crises of emergencies] *Investysii: praktyka ta dosvid*. – Kyiv. – May 2017. – №10. – S. 34–39. [in Ukrainian]
- [5] Krasnyuk M., Hrashchenko I., Krasniuk S., Kustarovskiy O. Reengineering of a Logistic Company and its Information System Taking into Account Macroeconomic Crisis. *Modern Economics*. 2019. Vol. 13(2019). pp. 141-153. DOI: [https://doi.org/10.31521/modecon.V13\(2019\)-23](https://doi.org/10.31521/modecon.V13(2019)-23).
- [6] Krasnyuk Maxim and Kustarovskiy Oleksandr. (2017) "The development of the concept and set of practical measures of anti-crisis logistics management in the current Ukraine conditions" // *Management theory & practice*. Publisher: Warsaw Management University. # 19 (1) 2017. pp. 31-38.
- [7] Krasnyuk, M., & Krasniuk, S. (2021). Association rules in finance management. *Збірник наукових праць ЛОГОΣ*. <https://doi.org/10.36074/logos-26.02.2021.v1.01>
- [8] Krasnyuk, M., & Krasniuk, S. (2020). Comparative characteristics of machine learning for predicative financial modelling. *Збірник наукових праць ЛОГОΣ*, 55-57. <https://doi.org/10.36074/26.06.2020.v1.21>
- [9] Krasnyuk, M., Tkalenko, A., & Krasniuk, S. (2021). Results of analysis of machine learning practice for training effective model of bankruptcy forecasting in emerging markets. // *Zu den materialien der internationalen wissenschaftlich - praktischen konferenz «Multidisziplinäre forschung: perspektiven, probleme und muster»* 9. april 2021. Wien, Republik Österreich <https://doi.org/10.36074/logos-09.04.2021.v1.07>
- [10] Kulynych Yu., Krasnyuk M., Tkalenko A., Krasniuk S. (2021). Methodology of Effective Application of Economic-Mathematical Modeling as the Key Component of the Multi-Crisis Adaptive Management. *Modern Economics*, 29(2021), 100-106. DOI: [https://doi.org/10.31521/modecon.V29\(2021\)-16](https://doi.org/10.31521/modecon.V29(2021)-16).