УДК 681.3.06 **doi: 10.26906/SUNZ.2020.4.059**

I. Ruban[1], I. Ilina[1], M. Mozhaiev[2]

[1] Kharkiv National University of Radio Electronics, Kharkiv, Ukraine
[2] Hon. Prof. M.S. Bokarius Kharkiv Research Institute of Forensic Examinations, Kharkiv, Ukraine

RESEARCHING PRIORITY DIRECTIONS IN THE AREA OF DATA MINING

Abstract. In the era of global informatization, social networks are acquiring great importance for obtaining various information by network users. But it must be borne in mind that social networks such as Facebook, Twitter, Instagram contain billions of raw unstructured data, the processing of which is indeed a rather difficult task for research. Data mining allows extracting current information from a large number of data sets, structuring and, after analyzing, gaining knowledge by detecting patterns among the data, which makes it possible to predict changes in the network that have occurred based on the interactions of information flows and events. This information is used in various areas such as business, education, medicine, cyber security, etc. The area of data mining has made tremendous success from its incipience to its current level, but Data Mining continues to face many challenges, especially when processing social media data. This article analyzes the various data mining methods that are used to analyze social networks, as well as explores the top priority areas in the field of data mining based on the review of various studies, and also focuses on the field of data mining in social networks, which will be used in further studies.

Ключові слова: Data Mining, social networks, data extraction, Data Mining methods and algorithms.

Introduction

Using large amounts of data is a hallmark of the 21st century, which produces amazing results when faced with another phenomenon of the century – social networks. Today, almost 96% of the world's population have access to social networks and this number has reached 2.34 billion people. Nowadays, out of the 100 most visited sites in the world there are 20 which are classic social networks and another 60 are socialized to one degree or another. More than 80% of companies around the world use social networks in their professional activities. About 78% of people trust information from social networks. By the number of users, Facebook is in first place – about 1.5 billion. Scientists have calculated that the minimum time a user spends on a social network is 3 hours, while he/she visits his/her account at least twice a day [1].

By registering on social networks, users are given the opportunity to communicate with relatives, friends, colleagues, and also make new acquaintances. Social networks can be used for self-development and self-study, gaining access to various information of interest. It is also possible to join a community on a specific topic and gain knowledge in specific areas. Social networks are a means for business development. Advertising can be directed to the target audience and one can find new customers, increase the loyalty of regular customers. Social networks have provided an opportunity to use a large amount of information, develop, improve and earn money.

But social networks have a number of negative features: due to the abundance of entertaining, superficial and often unnecessary garbage information, the time spent on the social network increases significantly. Such a pastime can negatively affect health, because a large amount of information often tires and burdens the nervous system. The disadvantage is that a person loses the skill of real communication, because he/she gets used to online communication. When texting on social networks, people often do not follow the rules of grammar and punctuation, use a poor vocabulary, emotions are replaced by emoticons – all this negatively affects communication in the real world. Intelligence agencies use social networks to organize surveillance. Intelligence agents account for approximately 15% of public network users.

It should also be taken into account bursts of group hate speech on social networks such as anti-Muslim, anti-immigrant, racist, anti-Semitic, homophobic, etc. A recent publication by the United Nations Educational, Scientific and Cultural Organization (UNESCO) indicates that "the nature of hate speech on the Internet and its impact on speech and offline behavior are poorly understood" and "the underlying reasons for these phenomena in which certain types of content appear, lead to actual discrimination, hostility or violence" [2].

Today, using social networks is continuously and rapidly increasing. Even more significant is the fact that these networks have evolved into a sizable pool of unstructured data belonging to multiple domains including business, governments, healthcare, etc. The data structure of social networks is not organized and is displayed in various forms such as text, voice, image and video [3]. Moreover, social networks provide a huge amount of continuous data in real time, including those in the form of time series, which makes traditional statistical methods not always effective for analyzing large amounts of data [4]. Consequently, data mining methods that combine mathematical and statistical approaches can play a significant role in overcoming this problem and facilitate data structuring. At the same time, the relevance of combined or hybrid forecasting methods, as well as methods for complexing forecast ratings [5], increases. It should also be taken into account that the primary data in many cases are characterized by various kinds of uncertainty, which necessitates the adaptation of traditional methods of analysis to such features [6].

Most of the predictive algorithms under these conditions are designed to function in real time [7].

Data Mining in Social Networks

When studying large-scale social networks data mining methods and algorithms are used. Analyzing the

data in social networks is of great interest for many reasons. For example, studying large social networks allows understanding social behavior in different contexts. In addition, by analyzing the roles of users involved in the network, it is possible to identify how information flows and opinions are disseminated in the network, and which of them are the most influential (Fig. 1). In addition, since the users of social networks may receive too much information from time to time, data mining of social networks can be used to support them by providing recommendations and filtering information on their behalf [4].

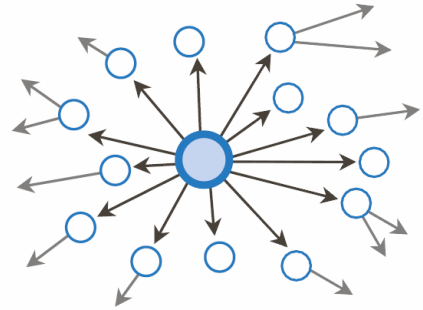

Fig. 1. Identification of the Most Influential Users in a Social Network

By understanding the characteristics of a particular social network, it is possible to build a mathematical model that explains the processes in the network. The mathematical model allows predicting future changes in the network, it becomes possible to simulate the behavior of users on the network.

Data mining includes the following stages: understanding and formulating the analysis problem, preparing data for automated analysis, applying Data Mining methods and building models, checking the constructed models, interpreting models by a person [8].

1. Collecting Data from Social Networks. Social networks, forums, news and entertainment portals and blogs contain a lot of material from which one can get information about the preferences and characteristics of users and companies. For this, first of all, it is necessary to identify the user in each source, which is not possible for all resources – on many of them users do not register or indicate insufficient data to identify them. Even where there is sufficient identification data, additional user information may not be useful. In this regard, social networks are the most suitable source containing information for identifying network participants, and additional data on preferences, marital status, education, social circle, etc.

In general, the task of enriching user profiles is as follows: using basic data (name, surname, date of birth, city), additional information is searched: circle of interests, social status, area of professional activity, musical preferences, etc. The task is to collect data about the client from social networks, identify him/her, enrich the data and form a single profile for each user (Fig. 2). One of the simplest ways to extract data is to use the services of specialized companies that collect and constantly update data from many sources. The main advantage here is the speed of obtaining information, which is essential with large volumes of the client base and the use of various social networks. The disadvantage is the paid subscription for data updates.

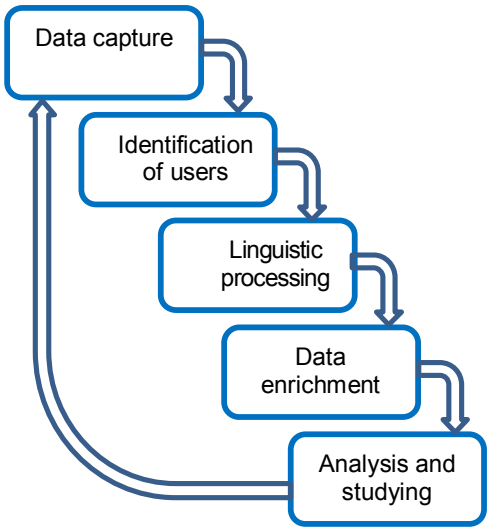

Fig. 2. Enriching User Profiles.

The next way is to use application programming interfaces (APIs) provided by almost all popular social networks. For different networks, APIs differ in the set of available data, restrictions on the number of requests, and the cost of accessing the interfaces. The disadvantages of this method include the limitation on the number of simultaneous requests and the number of calls that the application can make per unit of time. In addition, one need constantly to monitor changes in the APIs and update the data collection application, with some social networks providing important data only on a paid basis. The advantages of the method are the ability to retrieve data about a single user in a structured form (JSON or XML), as well as the ease of integrating API calls into one's own application.

One more way is the manual parsing of web pages of social networks, as well as the use of ready-made search robots (crawlers) to collect data with subsequent parsing. In this case, one has access to all open data and there are no restrictions on the speed of their collection. The disadvantages include the complexity of the implementation – the web page of each social network is unique, so each time one will have to develop one's own parsing rules, the complexity of support and the need for large computing resources, but this process is well parallelized.

The simplest identification method is to search for an exact match of all known user characteristics, but it must be borne in mind that the corresponding characteristics in social networks are reliable only to a certain extent – they may be absent or deliberately false, or allow different spellings. Therefore, before carrying out identification, it is necessary to cleanse and normalize the data, and also check the correctness of the parameters specified in the profile – for example, the user's city can be clarified based on the analysis of his subscriptions, posts and statuses. In addition to the data that netizens clearly indicate in their profiles, one can supplement the information by analyzing posts, subscription groups and photos.

In addition, there is unstructured content that is posted on social media – for example; messages (tweets), comments, audio or video. Automatic text analysis is impossible without the use of linguistic technologies. The correct combination of linguistic and statistical approaches increases the quality of the analysis result and the level of its reliability.

However, it is worth noting that mining data from social networks can involve serious legal and ethical issues, many of which are not legally regulated. Privacy issues are at the center of discussion concerning this tool. Regulating the use of data of social networks is necessary to protect freedom of expression among users of social networks. Companies using data of social networks often have their own views on how they will apply it. A striking example is the situation when Cambridge Analytica obtained access to the personal data of 87 million users of the social network Facebook without their knowledge in order to influence voters during the 2016 US presidential election [1]. Its data mining methods were contrary to Facebook policy. However, upon learning the hack, Facebook did not take serious legal action, which led to a scandal and the payment of a $ 5 billion fine for violating privacy.

2. **Methods for analyzing data in social networks.** Data mining allows automatically, based on a large amount of accumulated data, generating hypotheses that can be verified by other analysis tools. It is a computational process of identifying patterns or correlations in large relational databases using methods at the intersection of artificial intelligence, machine learning, statistics and database systems [9]. When analyzing 66 scientific publications [10] (filtered out of 1187) on this topic, the authors identified 19 data mining methods that have been used by researchers in the area of social networks over the past few years. Table 1 includes the detailed information on the frequency of occurrence of descriptions of data mining methods that have been encountered in the analyzed articles.

Table 1 – **Data Mining Methods Described in the Analyzed Publications and Their Frequency of Occurrence**

Method	Frequencies of Occurrence
AdaBoost	2
Artificial Neural Network (ANN)	8
Apriori	1
Bayesian Networks (BN)	26
Decision Trees (DT)	11
Density Based Algorithm (DBA)	3
Fuzzy Logic	1
Genetic Algorithm (GA)	1
Hierarchical Clustering (HC)	2
K Means	6
k of Nearest Neighbors (k NN)	9
Linear Discriminant Analysis (LDA)	9
Linear Regression (Lin R)	1
Logistic Regression (LR)	4
Markov Queueing Network	1
Maximum Entropy (ME)	2
New Methods	1
Support Vector Machines (SVM)	29
Wrapper	1

The diagram (Fig. 3) shows that the most frequently used methods in the field of social network analysis are (51% of analyzed articles) [10]: Support Vector Machines (SVM): it is considered an accurate classifier; resistant to noise; Bayesian Networks (BN): they allow improving classification by removing irrelevant features; they have good performance and low computing time; Decision Trees (DT): they give an accurate result; take up less memory; it takes less time to create a model; they have a short search time.

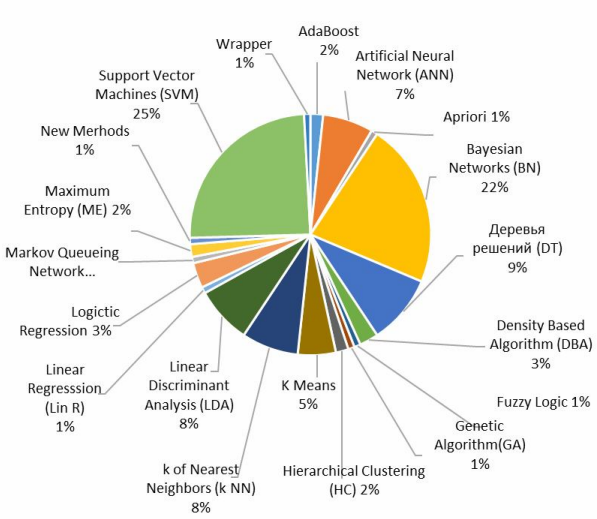

Fig. 3. Methods in the field of social network

The techniques that had a low frequency of occurrence, as each article was mainly devoted to its new method, was not taken into account.

Summarizing the data obtained, it can be carried out the following classification of data mining methods in social networks (for methods, the frequency of which was 3% and more) (Table 2).

The table shows that these methods were mainly used to analyze content, detect spam, assess users' preferences, detect inaccurate information, identify influential users, analyze users' characteristics, and only in some situations the methods were used for forecasting purposes (finance, medicine).

When working [10], the six areas of human activity were also identified, in which various research methods were used to analyze the flow of large data sets: business and management (BM); education (EDU); finance (FIN); government and the public (GP); medicine and health (MH); social networks (SN).

Fig. 4 shows that the most active areas that are used for data mining are: social networks, business and management accounting for 79%. The data analysis of social networks remains the most requested area of research. And recently, with an increase in the number of serious terrorist incidents in Western Europe and North America, the data analysis in social networks is aimed at identifying hate speech and predicting the reaction of users of social networks to these events.

Also, the nine active research tasks were identified, for the solution of which the methods of data mining are used [10]: biometric; content analysis; cybersecurity; disease awareness; geolocation; quality improvement; risk management.

Table 2 – **Classification Data Mining Methods**

Methods / Parameters	Artificial Neural Network (ANN)	Bayesian Networks (BN)	Decision Trees (DT)	Density Based Algorithm (DBA)	K Means	k of Nearest Neighbors (k NN)	Linear Discriminant Analysis (LDA)	Logistic Regression (LR)	Support Vector Machines (SVM)
Completing the task: classification; clustering	+	+	+	+	+	+	+	+	+
Approach to learning mathematical models: cybernetic; statistical	+	+	+	+	+	+	+	+	+
Solving forecasting problems	+	+	+			+		+	+
Social network content analysis	+	+	+		+		+	+	+
Analysis of characteristics of social network users		+		+	+	+			+
Spam filtering		+			+				+
Identifying the influential user	+								+

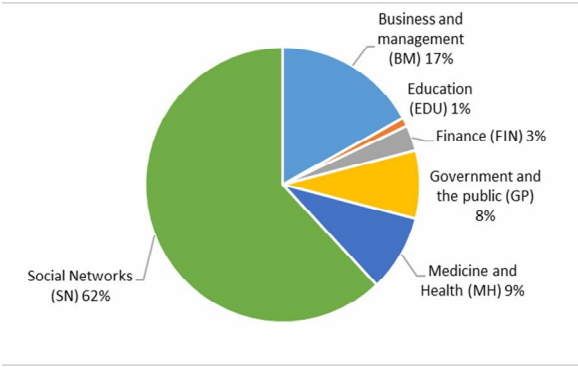

Fig. 4. Areas of Studying

Data mining methods are the process of extracting hidden knowledge from data. This can be done in a variety of ways such as decision trees, artificial neural network, Bayesian networks, k-NN, K-Means and SVM as machine learning methods. Also, statistical methods in some cases are considered as non-machine learning methods that are used to detect patterns [10-26]. As mentioned in [11], "statistical methods are data-driven and are used to discover patterns and build predictive models". Hence, machine learning and non-machine learning data mining methods such as traditional quantitative methods in statistics complement one another in data mining.

Conclusions

Based on the above, the following conclusions can be drawn: common data mining methods used in social networks are Support Vector Machines (SVM), Bayesian networks (BN) and Decision Trees (DT). SVM and BN are the most recommended methods for analyzing social networks, which are used and described in scientific articles [10]. Data mining methods have both advantages and disadvantages, and this makes the choice of certain methods dependent on the type of informative data required and can be quite a difficult task taking into account that the data transfer speed or data arrival is enormous and the dynamic nature of the data is also unpredictable.

In the area of social networks, more in-depth research is still needed that takes into account the precise application of data mining methods in the academic and industrial sectors. A careful study of the literature written in this area shows that a significant number of studies have used methods that solve segmentation problems, and to a lesser extent the methods that solve forecasting problems. Obviously, research in the area of social networks must include two main factors: sufficiently accurate and complete analyzed input data and the corresponding mining analysis to obtain predictive results.

REFERENCES

1. Carole Cadwalladr & Emma Graham-Harrison, Revealed: 50 Million Facebook Profiles Harvested for Cambridge Analytica in Major Data Breach, GUARDIAN (Mar. 3, 2018), https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-faceboo...
2. Gagliardone I., Gal, D., Alves, T. and Martinez, G.: Countering Online Hate Speech. UNESCO, Paris, France, 2015.
3. A.L. Kavanaugh, E. a. Fox, S.D. Sheetz, S. Yang, L.T. Li, D.J. Shoemaker, et al., Social media use by government: From the routine to the critical, Gov. Inf.Q.29, 2012. – P. 480-491.
4. H. Chen, R.H.L. Chiang, V.C. Storey, Business Intelligence and Analytics: From Big Data To Big Impact, Mis Q. 36, 2012.– P. 1165-1188.
5. Romanenkov, Yu. Formation of prognostic software support strategic decision-making in an organization [Text] / Yu. Romanenkov, V. Vartanian //Eastern-European Journal of Enterprise Technologies. – 2016. – Vol. 2, No. 9 (80). – PP. 25-34 (DOI: 10.15587/1729-4061.2016.66306).
6. Romanenkov, Yu. Complexification methods of interval forecast estimates in the problems on short-term prediction / Yu. Romanenkov, M. Danova, V. Kashcheyeva, O. Bugaienko, M. Volk, M. Karminska-Belobrova, O. Lobach // Eastern-European Journal of Enterprise Technologies. – 2018. – Vol. 3, No. 3 (93). – PP. 50-58 (DOI: 10.15587/1729-4061.2018.131939).

7. Romanenkov, Yu. Algorithmic Support for Auto-modes of adaptive short-term Forecasting in predictive Analytics Systems / Yu. Romanenkov, Yu. Pron-chakov, T. Zieiniiev // Proceedings of the XV International Scientific and Tech-nical Conference on «Computer Science and Information Technologies» (CSIT 2020). Volume II: Main Conference. Zbarazh-Lviv, Ukraine, 23-26 September, 2020. – P. 230-233.
8. Барсегян А.А. Куприянов М.С., Степаненко В.В., Холод И.И. Методы и модели анализа данных: OLAP и Data Mining. – СПб.: БХВ-Петербург, 2004. – 336 с.
9. Hemlata Sahu: A Brief Overview on Data Mining Survey, International Journal of Computer Technology and Electronics Engineering (IJCTEE) Volume 1, 2011. – P. 114-121.
10. MohammadNoor Injadat, Fadi Salo, Ali Bou Nassif: Data mining techniques in social media: A survey. Neurocomputing, Volume 214, 654-670 (2016).
11. S.Neelamegam: Classification algorithm in Data mining: An Overview, International Journal of P2P Network Trends and Technology (IJPTT), Volume 4, 2013. P.369-374.
12. Nechausov A., Mamusuĉ I., Kuchuk N. Synthesis of the air pollution level control system on the basis of hyperconvergent infrastructures. *Сучасні інформаційні системи*. 2017. Т. 1, № 2. С. 21-26. DOI: https://doi.org/10.20998/2522-9052.2017.2.04
13. Mozhaiev M., Kuchuk N., Usatenko M. (2019) The method of jitter determining in the telecommunication network of a computer system on a special software platform. *Innovative technologies and scientific solutions for industries*, 2019. Vol. 4 (10), pp. 134-140. doi: https://doi.org/10.30837/2522-9818.2019.10.134
14. Зиков І. С., Кучук Н. Г., Шматков С. І. Синтез архітектури комп'ютерної системи управління транзакціями e-learning. *Сучасні інформаційні системи*. 2018. Т. 2, № 3. С. 60–66. DOI: https://doi.org/10.20998/2522-9052.2018.3.10
15. Гахов Р.П. Моделирование трафика беспроводной сети передачи данных / Р. П. Гахов, Н. Г. Кучук// Научные ведомости БелГУ. – 2014. – № 1 (172). – Вып. 29(1). – С. 175-181.
16. Коваленко А. А., Кучук Г. А. Методи синтезу інформаційної та технічної структур системи управління об'єктом критичного застосування. *Сучасні інформаційні системи*. 2018. Т. 2, № 1. С. 22–27. DOI: https://doi.org/10.20998/2522-9052.2018.1.04
17. Свиридов А. С., Коваленко А. А., Кучук Г. А. Метод перерозподілу пропускної здатності критичної ділянки мережі на основі удосконалення ON/OFF-моделі трафіку. *Сучасні інформаційні системи*. 2018. Т. 2, № 2. С. 139–144. DOI: https://doi.org/10.20998/2522-9052.2018.2.24
18. Кучук Н. Г. Метод зменшення часу доступу до слабкоструктурованих даних / Н. Г. Кучук, В. Ю. Мерлак, В. В. Скородєлов // *Сучасні інформаційні системи = Advanced Information Systems*. – 2020. – Т. 4, № 1. – С. 97-102. doi: https://doi.org/10.20998/2522-9052.2020.1.14
19. Коваленко А.А. Использование временных шкал при аппроксимации длины очередей компьютерных сетей / А.А. Коваленко, Г.А. Кучук, И.В. Рубан // Сучасний стан наукових досліджень та технологій в промисловості. – 2018. – № 2 (4). – С. 12–18. – DOI: http://doi.org/10.30837/2522-9818.2018.4.012
20. Кучук Г.А. Минимизация загрузки каналов святи вычислительной сети / Г.А. Кучук // Системи обробки інформації. – Х.: НАНУ, ПАНМ, ХВУ, 1998. – Вип. 1(5). – С. 149-154.
21. Кучук Г. А., Можаєв О. О., Воробйов О. В. Метод агрегування фрактального трафіка. Радіоелектронні та комп'ютерні системи. 2006. № 6 (18). С. 181 - 188.
22. Donets V., Kuchuk N., Shmatkov S. Development of software of e-learning information system synthesis modeling process. *Сучасні інформаційні системи*. 2018. Т. 2, № 2. С. 117–121. DOI: https://doi.org/10.20998/2522-9052.2018.2.20.
23. Коваленко А. А. Подходы к синтезу информационной структуры системы управления объектом критического применения / А.А. Коваленко // Системи обробки інформації. – 2014. – № 1(117). – С. 180-184.
24. Raj Kumar: Classification Algorithms for Data Mining: A Surve, International Journal of Innovations in Engineering and Technology (IJIET) Vol. 1, Issue 2, 2012.– P.7-14.
25. Sagar S. Nikam: A Comparative Study of Classification Techniques in Data Mining Algorithms, Oriental journal of computer science & technology, Vol. 8, No. (1), 2015. – 13-19.
26. Nesma Settouti, Mohammed E, Amine Bechar and Mohammed Amine Chikh: Statistical Comparisons of the Top 10 Algorithms in Data Mining for Classification Task, International Journal of Interactive Multimedia and Artificial Intelligence, Vol. 4, No.1, 2016. – P.46-51.

Received (Надійшла) 22.09.2020
Accepted for publication (Прийнята до друку) 28.10.2020

Дослідження пріоритетних напрямів в області інтелектуального аналізу даних

І. В. Рубан, І. В. Ільїна, М. О. Можаєв

Анотація. В епоху глобальної інформатизації соціальні мережі набувають величезного значення для отримання різної інформації користувачами мереж. Але необхідно враховувати, що соціальні мережі такі як Facebook, Twitter, instagram містять мільярди необроблених неструктурованих даних, обробка яких дійсно є досить складним завданням для дослідження. Інтелектуальний аналіз даних дозволяє отримати поточну інформацію з великої кількості наборів даних, структурувати, і після проведеного аналізу отримати знання шляхом виявлення закономірностей між даними, що надає можливість прогнозування змін в мережі, які сталися на основі взаємодій інформаційних потоків та подій. Ця інформація застосовується в різних областях, таких як бізнес, освіта, медицина, кібербезпека і т.д. Область інтелектуального аналізу даних досягла величезних успіхів з моменту свого зародження до нинішнього рівня, але Data Mining продовжує стикається з багатьма проблемами, особливо при обробці даних соціальних мереж. Ця стаття присвячена аналізу різних методів інтелектуального аналізу даних, які використовуються для аналізу соціальних мереж, а також вивчення найбільш пріоритетних напрямків в області інтелектуального аналізу на основі проведеного огляду різних досліджень, а також фокусує увагу на області інтелектуального аналізу даних в соціальних мережах, що буде використано в подальших дослідженнях.

Ключові слова: Data Mining, соціальні мережі, вилучення даних, методи і алгоритми Data Mining.