ISSN 2226-3780

UDC 004.9
DOI: 10.15587/2312-8372.2018.123612

Lytvyn V.,
Vysotska V.,
Veres O.,
Brodyak O.,
Oryshchyn O.

BIG DATA ANALYTICS ONTOLOGY

Досліджені процеси аналізу Big Data. Використовуючи розроблену формальну модель та проведений критичний аналіз методів і технологій аналізу Big Data, побудовано онтологію аналізу Big Data. Досліджено методи, моделі та інструменти для удосконалення онтології аналітики Big Data та ефективнішої підтримки розроблення структурних елементів моделі системи підтримки прийняття рішень з керування Big Data.

Ключові слова: *онтологія аналізу Big Data, дані візуалізації, інтелектуальний аналіз даних, Text Mining, MapReduce.*

1. Introduction

With intensive business development, Big Data (BD) information technology (IT) is helping to preserve the competitiveness of the enterprise and handle significant volumes of accumulated structured and unstructured data. The application of methods and technologies for BD analysis and an integrated platform for Business Intelligence is relevant. BD allows to see and understand the links between pieces of information. This set of new tasks of public security, global economic models, privacy, established moral rules, legal relations of man, business and the state.

In connection with the rapid spread of smart and interconnected devices and systems, the volume of collected data is growing at an alarming rate. In some industries, about 90 % of data is stored in unstructured form, and their volume increases by more than 50 % annually. To maintain competitiveness, innovation and the rapid withdrawal of products and services to the market, it is necessary to be able to analyze these data and to obtain analytical information on their basis quickly and economically. With regard to BD analysis and other analytical tasks, current solutions do not provide the information system (IS) response speed necessary for working with analysis tasks, reduces user productivity and delays the decision making process [1, 2]. Consumers are changing, the world of business is changing. Today it is not enough to study only the data on sales. The goal of deploying an integrated platform for Business Intelligence (BI) and BD analysis is to dig deeper and better understand why, where, what and how – about customers, products and the company. The methods of doing business are changing. The behavior of consumers is changing. The consumers themselves are changing. In order to remain competitive, businesses are seeking in real time to find out when customers buy something, they buy, and even what they think before entering a store or visiting a Web site. BD, BD analysis and an integrated platform for BI and BD analysis help in this [1–4].

2. The object of research and its technological audit

The object of this research is the Big Data analysis processes.

At the input of the system are methods and IT of BD analysis, which are described in [1–5]. At the output of the system is the ontological model of the BD analysis rules $O=<X, R, F>$. The taxonomy of ontology X defines the Big Data A_{BD} analysis technique (root concept of ontology). The optimal definition of the set of relations between these concepts of R and the set of rules F of the BD analysis, formalized with the help of descriptive logic DL, will effectively handle BD, that is: S: RABD $\rightarrow O$.

One of the most problematic places is the lack of a clear classification of BD analysis methods, the presence of which will greatly facilitate the selection of an optimal and efficient algorithm for analyzing these data depending on their structure. BD analysis is of great practical importance for modern IT and serves the solution of actual everyday problems, but it generates even more new ones. Effective and timely BD analysis is able to change our way of life, work and thinking. One of the conditions for the successful development of the world economy at the present stage is the ability to capture and analyze huge arrays and information flows. Countries that will master the most effective methods of working with BD, are waiting for a new industrial revolution. The direction of «Big Data» concentrates efforts in organizing the storage, processing, analysis of huge data sets. A common mistake around large amounts of data is the assumption that the acquisition of a powerful computer infrastructure will immediately provide benefits to the business, instead of IT, informatics and mathematics going hand in hand. Infrastructure is necessary, but benefiting from large amounts of data also requires more sophisticated methods of analyzing them.

3. The aim and objectives of research

The aim of research is development of a software system S to formalize the Big Data RABD analysis rules in the form of an ontological knowledge base (KB) for its use for processing and analyzing any BD.

To achieve this aim it is necessary to:

1. Investigate the features of the classification of methods and technologies of Big Data analytics, taking into account the definition and application of the relevant IT.

2. Develop a formal BD analysis model.
3. Develop an ontological KB of BD analysis.
4. Build the Big Data RABD analysis rules.

4. Research of existing solutions of the problem

Standard business practice for large-scale data analysis is based on the concept of EDW (Enterprise Data Warehouse), requests to which come from BI software [1–5]. BI tools allow to create reports and interactive interfaces, to aggregate data using aggregate functions in different hierarchical data distributions to groups.

A carefully designed EDW plays a central role in the proper application of IT. The design and evolution of a detailed knowledge warehouse (KW) schema is a general principle of disciplined integration of the data of large enterprises, improving the results and presentation of all business processes. The resulting database (DB) plays the role of a repository of characteristics of critical business functions. In addition, the database server, which stores the KW, is traditionally the main computing facility, serves as a central, scalable mechanism for key corporate analytics. The conceptual and computational central position of the KW makes it a critically important resource that is used to generate reports with a large amount of data. And these reports target decision-makers. The KW is traditionally controlled by specially appointed IT staff, who not only escort the IS, but also carefully control access to it so that management can guarantee a high level of service [5].

Although in many situations this orthodox KW approach continues to be applied, a number of factors contribute to a completely different philosophy of managing large-scale data in enterprises. Data storage is now so cheap that small sub-groups of the enterprise can develop a separate astronomical database within their own budget. The number of intracorporate large-scale data sources is growing significantly: large databases now appear even on the basis of a single source of click-stream data streams, IP logs, e-mail archives and forums, and the like. The importance of data analysis becomes widely recognized. Numerous companies demonstrate that sophisticated data analysis helps to reduce costs and even direct revenue growth. The result of these opportunities is a massive shift to the collection and use of data in several organizational units of corporations. The advantage of this transition is increasing the efficiency and growth of the culture of data use, but it enhances the data decentralization, which KW is called upon to combat. In this changing climate of collecting disparate large-scale data, the MAD approach (Magnetic, Agile, Deep data analysis) [5] is appropriate.

In modern analysis, BD uses increasingly sophisticated statistical methods that go far beyond the generalization (rollup) and drilldown of traditional BI methods. When executing these algorithms, analysts often need to explore huge sets of data without resorting to the use of samples. Modern KW should serve as a solid (deep) data repository, and a mechanism to support the implementation of complex algorithms. Today, there is a growing need for powerful data analysts. Often they are highly qualified statisticians with a good knowledge of software, but usually focus on thorough data analysis, rather than on database management. To support their activities, it is required to apply the MAD approach to the KW design and the creation of the infrastructure of the database systems. When these goals are achieved, important problems arise in the choice of methods and IT for BD analysis. Working with BD is not like a normal BI process, where simply adding known values brings a result. When working with BD, the result is obtained in the process of cleaning them by sequential modeling: first a hypothesis is put forward, a statistical, visual or semantic model is constructed, on the basis of this, the reliability of the hypothesis is checked and the following is advanced. This process requires the researcher or interpreting visual values, or compiling interactive knowledge-based queries, or developing adaptive ML algorithms that can produce the desired result. And the lifetime of such algorithm is often quite short [2, 6]. There are five basic approaches to analyzing BD [7]:

1. *Discovery tools* are useful during the information lifecycle for a quick, intuitive study and analysis of information derived from any combination of structured and unstructured sources. These applications allow analysis of data sources along with traditional BI systems. There is no preliminary modeling, users quickly attract new ideas, form meaningful conclusions, and make informed decisions.

2. *BI tools* are essential for reporting, analyzing and managing performance, primarily with transactional data from SD and IS production. Applications provide great opportunities for BI and performance management.

3. *In-DatabaseAnalytics* – methods for finding patterns and relationships in data. Applied to the database, there is no transfer of data from other analytical servers, speeds up the information processing cycle and reduces the total cost.

4. *Hadoop* – preliminary processing of data for trends of macro identity or finding of data elements the value of the OUTOF-range. Organizations use Hadoop as a precursor for forms of analytics.

5. *Management solutions* – predictive modeling, business rules and self-learning to make an informed decision based on the current context. Creates decision-making processes in real time.

All these approaches are used to identify hidden relationships.

Today, there is no difference in the use of the terms Big Data and Big Data Analytics. These terms describe both the data itself and the control technologies and analysis methods [8–10]. Big Data Analytics is the development of the Data Mining concept. The same are the tasks, fields of application, data sources, methods and IT. Since the advent of the Data Mining concept, before the advent of the BD era, the volumes of analyzed data have changed in a revolutionary way, there appeared IP high-performance computing, new IT, including MapReduce and its numerous software. With the advent of social networks, new tasks have emerged. Data Mining is a process of decision support based on the search for raw data in hidden patterns, previously unknown, non-trivial, practically useful and accessible interpretation of knowledge necessary for decision-making in various spheres of human activity [10–12].

Data Mining is an approach to data analysis. The emphasis is not only on the extraction of facts, but also on the generation of hypotheses. The hypotheses created in the process should be checked with the help of the usual analysis within the framework of the usual schemes and/or with the help of experts of the subject area (software).

ISSN 2226-3780

This approach uses traditional analysis tools, such as mathematical statistics (regression, correlation, cluster, factor analysis, time series analysis, decision trees, etc.). And also those tools related to artificial intelligence (AI) (ML, neural networks, genetic algorithms, fuzzy logic, etc.). If the DataMining approach adds MapReduce technology and the 4V requirement (Volume, Velocity, Variety, Veracity), then this will show the functional links of Big Data Analytics. Analysis of large amounts of data and the need to understand the meaning of individual behavior require processing methods that go beyond traditional statistical methods [10–13]. In [13], a draft list of methods and methods of BD analysis, which does not claim to be complete, but it reflects the approaches most demanded in various branches. In addition, some of the BD data can be successfully used for smaller arrays (for example, A/B testing, regression analysis). Undoubtedly, the larger and diversified array is amenable to analysis, the more accurate and relevant data can be obtained at the output.

5. Methods of research

Big Data is a series of approaches, tools and methods for processing structured and unstructured data of huge volumes. It is also a source of considerable diversity for obtaining understandable results that are effective in the conditions of continuous growth, distribution by network nodes, alternative to traditional database management systems and BI class solutions [7]. There are three types of tasks related to BD [1–4, 6, 7]: storage and management, processing of unstructured information, analysis of BD (Fig. 1).

The formal BD model as IT has this representation [8, 9]:

$$BD=\langle Vol_{BD},\ Ip,\ A_{BD},\ T_{BD}\rangle, \quad (1)$$

where Vol_{BD} – the set of types of volumes; Ip – a set of types of data sources (information products); A_{BD} – set of methods of Big Data analysis; T_{BD} – set of technologies for Big Data processing.

Proceeding from the definition of BD [9], it is possible to formulate the basic principles of working with such data: horizontal scalability, resistance to failures, and locality of data. All modern means of working with BD in one way or another correspond to these three principles. In order to comply with them, it is necessary to come up with some methods and paradigms for developing data processing tools. Today, there are a lot of $A_{BD}=\{A_i\}$ different methods of analyzing data arrays, based on tools borrowed from statistics and informatics (Fig. 2, 3).

Groups of methods and technologies for BD analysis can be formally represented in the form of a tuple:

$$A_{BD}=\langle M_{Data\ Mining},\ M_{Machine\ Learning},\ M_{Visualization},\ T_{Text\ Mining},\ T_{MapReduce},\ T_{other},\ K_{BD},\ fdm,\ fml,\ fmv,\ fmt,\ fmr,\ fmo\rangle, \quad (2)$$

where $M_{Data\ Mining}$ – a set of Data Mining methods; $M_{Machine\ Learning}$ – a set of Machine Learning methods; $M_{Visualization}$ – methods for graphical representation of BD analysis; $T_{Text\ Mining}$ – Text Mining technologies; $T_{MapReduce}$ – MapReduce technologies; T_{other} – other specific methods and technologies for BD analysis; fdm – the function of determining the Data Mining method in accordance with the type of task; fml – the function of determining the Machine Learning in accordance with the type of task; fmv – the function of determining the methodology for graphical representation of the BD analysis in accordance with the type of task; fmt – the function of determining the Text Mining technology in accordance with the type of task; fmr – function of determining the MapReduce technology according to type of task; fmo – function of determining another BD analysis technology in accordance with the type of task.

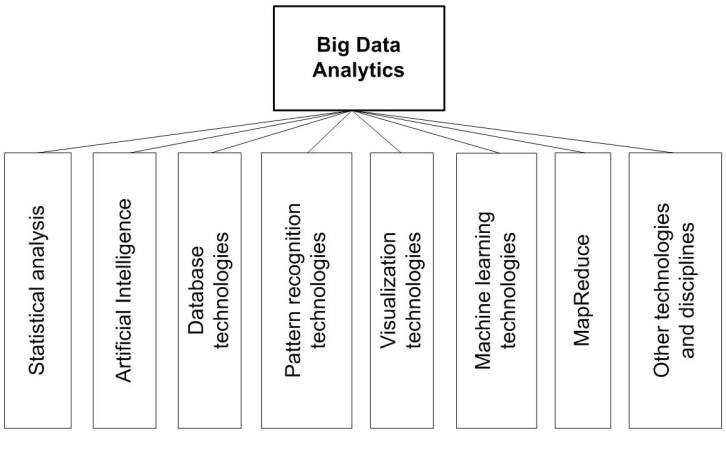

Fig. 1. Functional relations of Big Data analytics

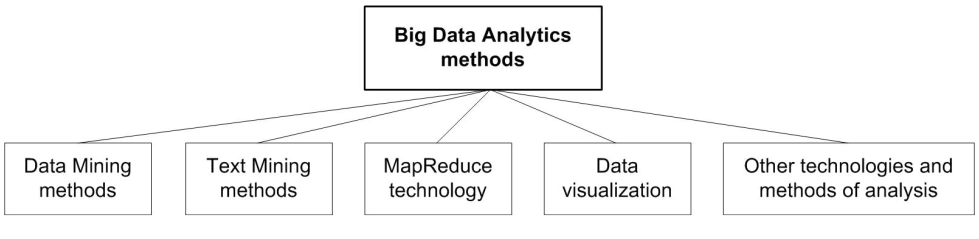

Fig. 2. Groups of methods of Big Data analytics

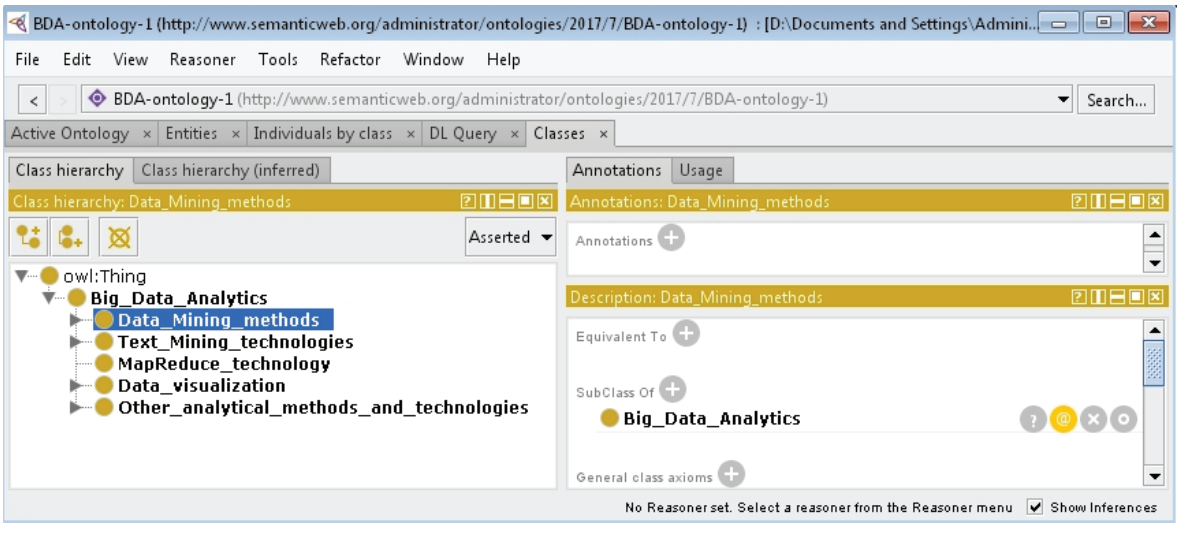

Fig. 3. Subclasses of the Big Data analysis class using Protege 3.4.7

And

$$K_U \subset K_{BD}$$

and

$$K_{BD}=K_{Data\ Mining} \cup K_{Machine\ Learning} \cup \cup K_{Visualization} \cup K_{Text\ Mining} \cup K_{MapReduce} \cup K_{other}, \quad (3)$$

where K_{BD} – criteria and parameters of the BD analysis; K_U – criteria and parameters for the analysis of a specific BD; $K_{Data\ Mining}$ – criteria and parameters for choosing the Data Mining method in accordance with K_U; $K_{Machine\ Learning}$ – criteria and parameters of the choice of the method of machine learning in accordance with K_U; $K_{Visualization}$ – criteria and parameters of the choice of the method of visualization in accordance with K_U; $K_{Text\ Mining}$ – criteria and parameters for choosing the Text Mining technology in accordance with K_U; $K_{MapReduce}$ – criteria and parameters for choosing the MapReduce technology in accordance with K_U; K_{other} – criteria and parameters for choosing another BD analysis method, respectively, in accordance with K_U.

The need for new tools for BD analysis is justified by the fact that the data becomes larger, larger than their external and internal sources, now they are more complex and diverse (structured, unstructured and poorly structured), different indexing schemes are used (relational, multidimensional, noSQL). The previous methods of data processing are inefficient – Big Data Analytics extends to large and complex arrays, including Discovery Analytics and Exploratory Analytics. Data Mining Data mining is the manifestation of hidden relationships or patterns between variables in large amounts of raw data. The choice of the Data Mining method for BD analysis depends on the type of task. Accordingly, according to (2) MData Mining is represented as a tuple:

$$M_{Data\ Mining}=\langle Tk_{Data\ Mining},\ Md_{Data\ Mining},\ fdm,\ K_{Data\ Mining}\rangle, \quad (4)$$

where $Tk_{Data\ Mining}$ – Data mining tasks at $Tk_{Data\ Mining}=fdm(Md_{Data\ Mining}, K_{Data\ Mining})$; $Md_{Data\ Mining}$ – Data mining methods.

The use of Data Mining methods allows to solve the following problems [14–18]:

$$Tk_{Data\ Mining}=\langle T_{Classification},\ T_{Clustering},\ T_{Associations},\ T_{Sequence},\ T_{Forecasting},\ T_{Deviation\ Detection},\ T_{Estimation},\ T_{LinkAnalysis},\ T_{Graph\ Mining},\ T_{Summarization}\rangle, \quad (5)$$

where $T_{Classification}$ – identification of characteristics describing groups of objects of sets of data under study – classes; according to these characteristics, the new object will belong to one or another class; $T_{Clustering}$ – clustering (separation) of objects into groups; $T_{Associations}$ – finding the patterns between the related events in the data set; $T_{Sequence}$ – identification of the relationship between time-related events (the sequence is characterized by a high probability of a chain of time-related events); $T_{Forecasting}$ – on the basis of special properties of accumulated data, future values of indicators are estimated; $T_{Deviation\ Detection}$ – identification and analysis of data, most differ from the total number of data, the identification of uncharacteristic patterns; $T_{Estimation}$ – forecast of continuous values of characteristics; $T_{LinkAnalysis}$ – finding dependencies in the data set; $T_{Graph\ Mining}$ – the creation of a graphic image of the analyzed data to illustrate the existence of regularities in the data; $T_{Summarization}$ – description of specific groups of objects using the data set in question.

So, according to (3), (4) for the solution $T_{Classification}=ftc(MT_{Data\ Mining})$ use:

$$MT_{Data\ Mining}=\langle M_{Nearest\ Neighbor},\ M_{k\text{-}Nearest\ Neighbor},\ M_{Bayesian\ Networks},\ M_{Tree},\ M_{Neural\ Networks}\rangle, \quad (6)$$

where $M_{Nearest\ Neighbor}$ – nearest neighbors method for data classification; $M_{k\ Nearest\ Neighbor}$ – Nearest Neighbor method for data classification; $M_{Bayesian\ Networks}$ – Bayesian Networks for data classification; $M_{Neural\ Networks}$ – Neural Network for data classification; M_{Tree} – induction of decision trees for data classification; ftc – function of defining the Data Mining method for the classification problem.

The most famous algorithm for solving $T_{Associations}=apriori(Data,\ Signs,\ Rules)$.

ISSN 2226-3780

Data Mining is a set of techniques that allows you to determine the most receptive to the promoted or service category of consumers, identify the characteristics of the most successful employees, provide for the behavioral model of consumers, etc. [10–12], that is:

$$Md_{Data\ Mining} = \langle MD_{Supervised\ Learning}, MD_{Unsupervised\ Learning}, MD_{St}, MD_{Cb} \rangle, \quad (7)$$

where $MT_{Data\ Mining}$ – set of Data Mining methods for the classification problem; $MD_{SLearning}$ – set of Data Mining methods for Supervised Learning; $MD_{ULearning}$ – set of Data Mining methods for Unsupervised Learning; *MDSt* – statistical Data Mining methods for database analysis; *MDCb* – cybernetic Data Mining methods for database analysis.

Another classification of Data Mining methods is based on different approaches to teaching mathematical models (Fig. 4, 5) [14–18].

Statistical methods of Data Mining contain: preliminary analysis of the nature of statistical data, the identification of relationships and regularities, multivariate statistical analysis, dynamic models and forecast based on time series:

$$MD_{St} = \langle MS_1, MS_2, MS_3, MS_4 \rangle, \quad (8)$$

where MS_1 – descriptive analysis and description of the source data; MS_2 – correlation analysis (correlation and regression, factor, variance); MS_3 – multidimensional statistical analysis (component, discriminant, multivariate regression, canonical correlations; MS_4 – time series analysis (dynamic models and forecasting).

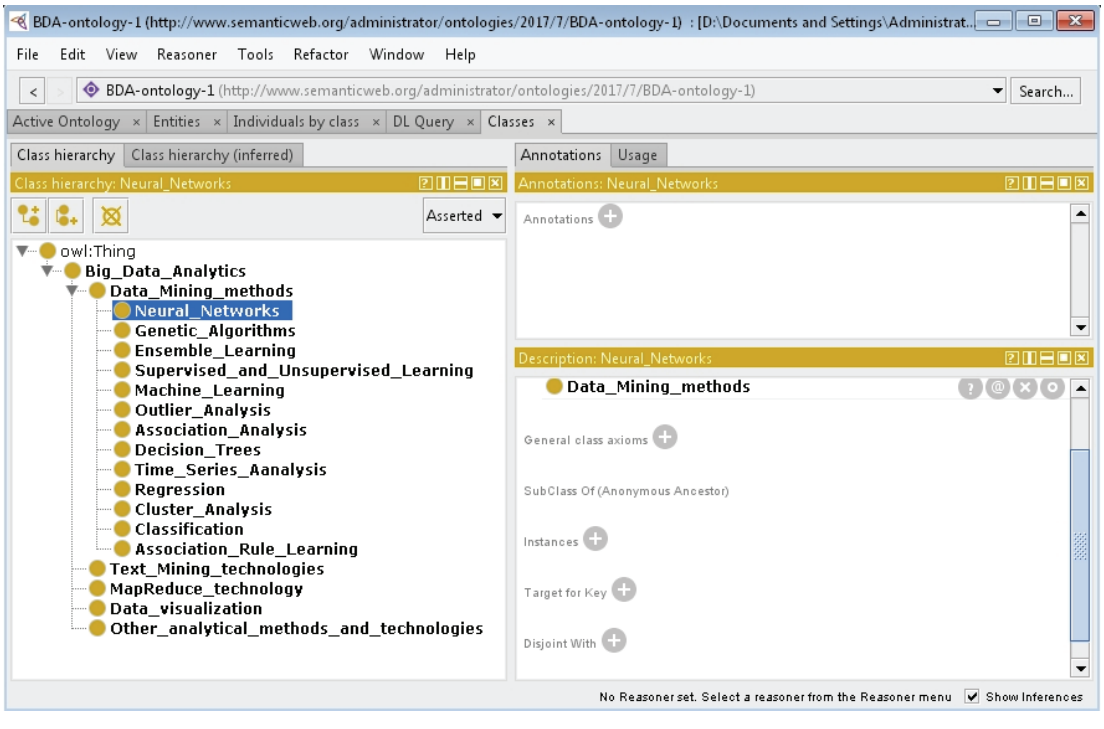

Fig. 4. Subclasses of Data Mining Methods class

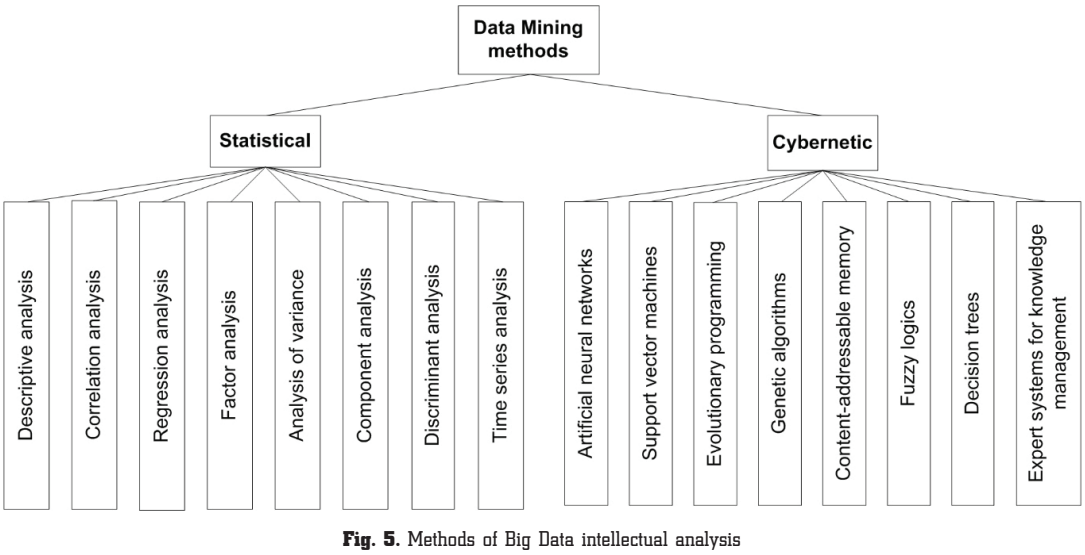

Fig. 5. Methods of Big Data intellectual analysis

To solve the clustering problem on graphs, Girvanand Newman algorithm, of the MLP method (Markov Cluster Algorithm) is to be implemented. An example is the segmentation of the market. Regression analysis is used to determine:

– the impact of customer satisfaction on customer loyalty;

– dependence of the number of support for received calls from the weather forecast, taking into account the previous day;

– the influence of neighborhood and size on the price of listing houses;

– compatibility in the life of the user through the online dating site and the like.

Analysis of time series – tracking the securities market or the incidence of patients. Emission analysis is used to identify fraud, personal marketing, medical analysis [19–23]. The cybernetic methods of DataMining include methods [23–27]:

$$MD_{Cb}=\langle MC_1,\ MC_2,\ MC_3,\ MC_4,\ MC_5,\ MC_6,\ MC_7\rangle, \quad (9)$$

where MC_1 – evolutionary programming; MC_2 – associative memory (search for analogs, prototypes); MC_3 – fuzzy logic; MC_4 – decision trees; MC_5 – expert knowledge processing systems; MC_6 – artificial neural networks (recognition, clustering, prediction); MC_7 – genetic algorithms (optimization).

MC_6 is a class of models based on the analogy with the work of the human brain and is intended for solving various problems of data analysis after passing through the learning phase on data. MC_6 is a model of the biological neural networks of the brain, in which neurons are simulated by the same type of elements (artificial neurons). MC_6 is used to solve the following problems:

– automation of image recognition processes;

– forecasting the performance of the enterprise;

– medical diagnostics; forecasting;

– adaptive management;

– creation of expert systems;

– organization of associative memory;

– processing of analog and digital signals;

– synthesis and identification of electronic systems.

With the help of MC_6, it is possible, for example, provide sales volumes, market indicators, recognize signals, develop self-learning systems.

MC_7 inspired by the nature of evolutionary processes, that is, through mechanisms such as inheritance, mutation and natural selection. These mechanisms are used to evolve a useful solution to problems that require optimization. MC_7 is used for the following tasks:

– scheduling physicians for the emergency hospital;

– creation of a combination of optimal materials and engineering methods necessary for the development of economical vehicles;

– generation of artificially creative content, such as puns and jokes;

– forecasting the stock market index using time series analysis. To analyze the market basket apply the analysis of hidden patterns (Association Analysis). Statistical classification is used to automatically assign a category to documents; classification of organisms into groups; separate profiles of students who take online courses; purposeful selection of employees (focuse dhiring), etc.

Another interesting area in artificial intelligence (AI) for BD analysis is Machine Learning (ML). This direction in computer science (historically behind it was named «artificial intelligence»), which aims to create self-learning algorithms based on the analysis of empirical data. ML is formed as a result of the separation of the science of neural networks into methods of learning networks and types of topologies of the architecture of networks. It also contains methods of mathematical statistics. The following ML methods are based on neural networks. The basic types of neural networks, namely, perceptron and multilayer perceptron (and their modifications) can be trained as a teacher, without a teacher, with reinforcement and actively. However, most statistical methods and some neural networks belong to only one of the ways of learning:

$$M_{Machine\ Learning}=\langle ML_1,\ ML_2,\ ML_3,\ ML_4,\ ML_5,$$
$$ML_6,\ ML_7,\ ML_8,\ ML_9,\ ML_{10},\ fml,\ K_{Machine\ Learning}\rangle, \quad (10)$$

where ML_1 – methods of supervising learning; ML_2 – methods of unsupervising learning; ML_3 – learning methods with reinforcement; ML_4 – methods of active learning; ML_5 – learning methods with partial involvement of the teacher; ML_6 – methods of transductive learning; ML_7 – multitasking learning methods; ML_8 – methods of diverse learning; ML_9 – methods of controlled and uncontrolled learning; ML_{10} – methods of the learning ensemble.

ML contains software that is able to extract knowledge from the database. This gives the IS the opportunity to learn without being explicitly programmed, and focuses on prediction based on known properties extracted from the learning data sets. Today ML use:

– to detect spam and non-spam e-mail messages;

– to gain knowledge about user preferences and recommendations based on this information;

– to determine the best content for attracting potential customers;

– to obtain the probability of winning the case and establishing the legal norms of the accounts presented.

ML_9 allow to reveal functional interrelations in the considered data sets. ML_{10} uses a lot of predicative models, which improves the quality of the predictions made.

The methods for graphically representing the results of BD *MVisualization* analysis in the form of diagrams or animations greatly simplify the interpretation and make it easier to understand the obtained results [27–31]. New progressive methods of visualization are:

$$M_{Visualization}=\langle MV_1,\ MV_2,\ MV_3,\ MV_4,$$
$$fmv,\ K_{Visualization}\rangle, \quad (11)$$

where MV_1 – teg cloud; MV_2 – clusterogram; MV_3 – historical stream; MV_4 – the spatial flow.

Each element in MV_1 is assigned a specific weighting factor that correlates with the font size [32–36]. In the case of text analysis, the value of the weight coefficient directly depends on the frequency of use (citation) of a specific word or phrase and allows the reader in a short time to get an idea of the key points of an arbitrarily large text or set of texts. MV_2 shows how individual elements of a set of data correlate with clusters as their number changes. Choosing the optimal number of clusters is an

ISSN 2226-3780

important part of cluster analysis. MV_3 helps to monitor the evolution of the document, the creation of which works simultaneously a large number of authors. The horizontal axis lays time, on the vertical axis – the contribution of each of the co-authors, that is, the amount of text entered. Each unique author is assigned a certain color in the diagram. MV_4 allows to track the spatial distribution of information. The brighter the line, the more data is transferred per unit of time [36, 37].

The basis of $T_{Text\ Mining}$ is statistical and linguistic analysis, methods of artificial intelligence. This technology is used for analysis, navigation and search in unstructured texts [38–42]. Using IT class $T_{Text\ Mining}$ allows users to acquire new knowledge. This is a set of methods that are designed to extract information in texts based on modern IT, allows to identify patterns and ensure that users receive useful data and new knowledge (Fig. 6):

$$T_{Text\ Mining}=\langle MT_1,\ MT_2,\ MT_3,\ MT_4,\ MT_5,\ MT_6,\ MT_7,\ MT_8,\ MT_9,\ MT_{10},\ MT_{11},\ fmt,\ K_{Text\ Mining}\rangle, \quad (12)$$

where MT_1 – *classification*; MT_2 – *clustering*; MT_3 – Building Semantic Networks or Analyzing Relationships (*Relationship, Event andFact Extraction*); MT_4 – extraction of phenomena, facts, concepts (feature extraction); MT_5 – automatic abstracting, creating annotations (summarization); MT_6 – question answering; MT_7 – thematic indexing; MT_8 – keyword searching; MT_9 – means of support and creation of taxonomy (oftaxonomies); MT_{10} – means of support and creation of thesauri (thesauri); MT_{11} – methods and means of content analysis (Content Analysis).

$T_{Text\ Mining}$, like most cognitive technologies, is an algorithmic identification of previously unknown links and correlations in already available text data. $T_{Text\ Mining}$ widely uses the methodology and approaches of data mining technology, for example, MT_1 or MT_2. $T_{Text\ Mining}$ has new capabilities: automatic abstracting of texts and the identification of phenomena, that is, concepts and facts. An important task of $T_{Text\ Mining}$ is extraction from the text of its characteristic elements or properties, which can be used as document metadata, keywords, annotations. Another important task is establishment of the document attribution to certain categories from a given scheme of their systematization. $T_{Text\ Mining}$ provides a new level of semantic search for documents. $T_{Text\ Mining}$ features are used to solve the task of identifying templates in the text, automatically «pushing» or distributing data by profiles, creating document reviews.

This is a tool that enables to analyze BD in search of trends, patterns and relationships that can help in making strategic decisions. The main purpose of $T_{Text\ Mining}$ is giving the analyst the opportunity to work with BD by automating the process of obtaining the right data. As an example of effective application of $T_{Text\ Mining}$ technologies is MT_{11}, which is characterized by the objectivity of the conclusions and the rigor of the procedure. Its basis is the quantification of the text with subsequent interpretation of the results. The subject of MT_{11} can be both problems of social reality, which are expressed or vice versa hiding in documents, and the internal laws of the object of research itself [38]. The popularity of MT_{11} is based on the fact that this method allows measuring human behavior (assuming that verbal behavior is its form). Unlike surveys, content analysis measures not what people say they did or will do, but they really did.

Let's describe several technologies and disciplines of data research from the point of view of BD technology for T_{other} (Fig. 7) [14–19]:

$$T_{other}=\langle MO_1,\ MO_2,\ MO_3,\ MO_4,\ MO_5,\ MO_6,\ MO_7,\ MO_8,\ MO_9,\ MO_{10},\ MO_{11},\ MO_{12},\ MO_{13},\ fmo,\ K_{other}\rangle, \quad (13)$$

where MO_1 – метА/В testing, Splittesting methods; MO_2 – Natural Language Processing, NLP methods; MO_3 – Sentiment Analysis methods; MO_4 – Network Analysis methods; MO_5 – Optimization methods; MO_6 – Pattern Recognition methods; MO_7 – Predictive Modeling methods; MO_8 – Signal Processing methods; MO_9 – Spatial Analysis methods; MO_{10} – Statistics methods; MO_{11} – Simulation methods; MO_{12} – Crowdsourcing methods; MO_{13} – Data Fusion and Data Integration methods.

MO_1 is used when optimizing Web pages in accordance with a specified goal. MO_3 is based on MO_2. They allow to extract messages from the general information flow associated with an interested item (for example, a consumer product). Next, evaluate the polarity of the judgment (positive or negative), the degree of emotionality, and the like. MO_3 helps researchers determine the mood of speakers or authors in relation to the topic. The analysis of moods is used to help: improve the quality of service in the hotel network, analyzing guest comments; set up incentives and services to address what customers are really asking for; determine which consumers are really under the influence of social media. MO_4 is a technique for analyzing connections between nodes in networks. With respect to social networks, it is possible to analyze the relationship between individual users, companies, communities, and the like. MO_5 is designed for the redesign of complex systems and processes to improve one or more indicators. Helps in making strategic decisions, for example, the composition of the product line being introduced to the market, conducting investment analysis, and the like. MO_7 allow to create a mathematical model of a pre-defined probable scenario of events. For example, analysis of the database CRM-system for possible conditions that will push subscribers to change the provider. MO_{12} – categorization and enrichment of data by a broad, indefinite circle of people, with the aim of using their creative abilities, knowledge and experience in applying information and communication technologies. MO_{13} allows to analyze the comments of users of social networks and compare with the results of sales in real time.

$T_{MapReduce}$ is a distributed computing model introduced by Google, which is used for parallel computations over very large (several petabytes) data sets in computer clusters [42–47]. In terms of implementation, the analytical platform for working with BD should be able to use the new $T_{MapReduce}$. In practice, BD analysis is rarely to calculate statistical totals for all data. Instead, the importance of BD lies in the ability to separate data into micro-segments and use the methods of intellectual analysis and predictive modeling to build a large number of models for small groups of observations. There are many tools for performing such aggregation of data in a distributed file system, which makes it easy to implement this analytical process.

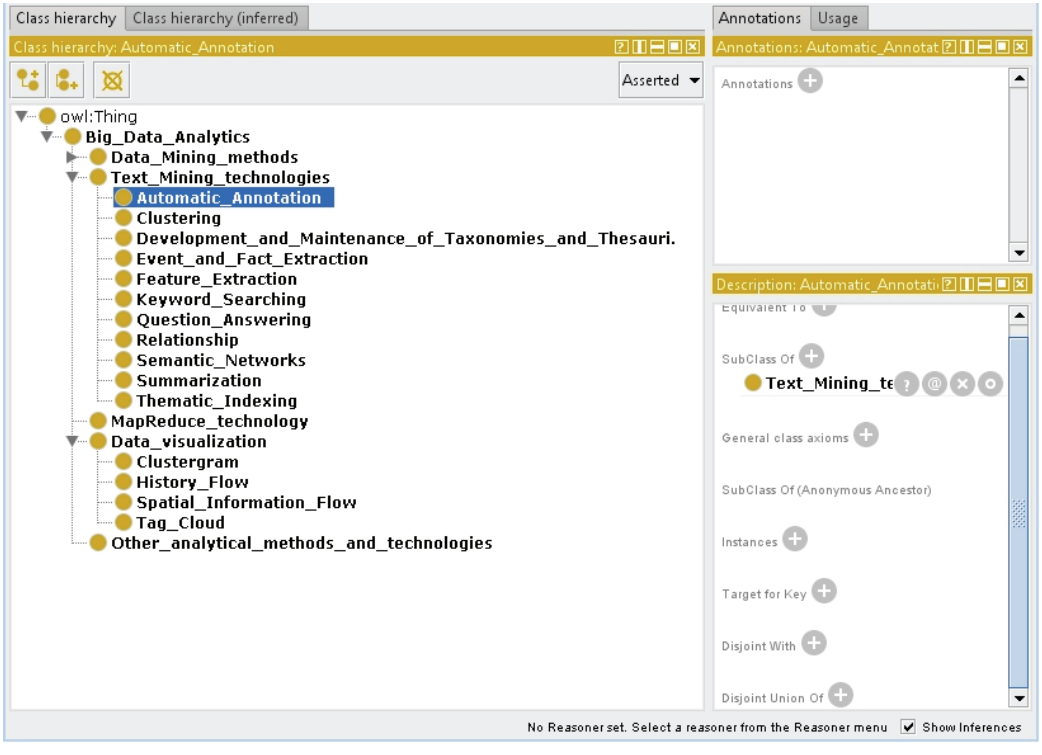

Fig. 6. Subclasses of Text Mining Technologies class

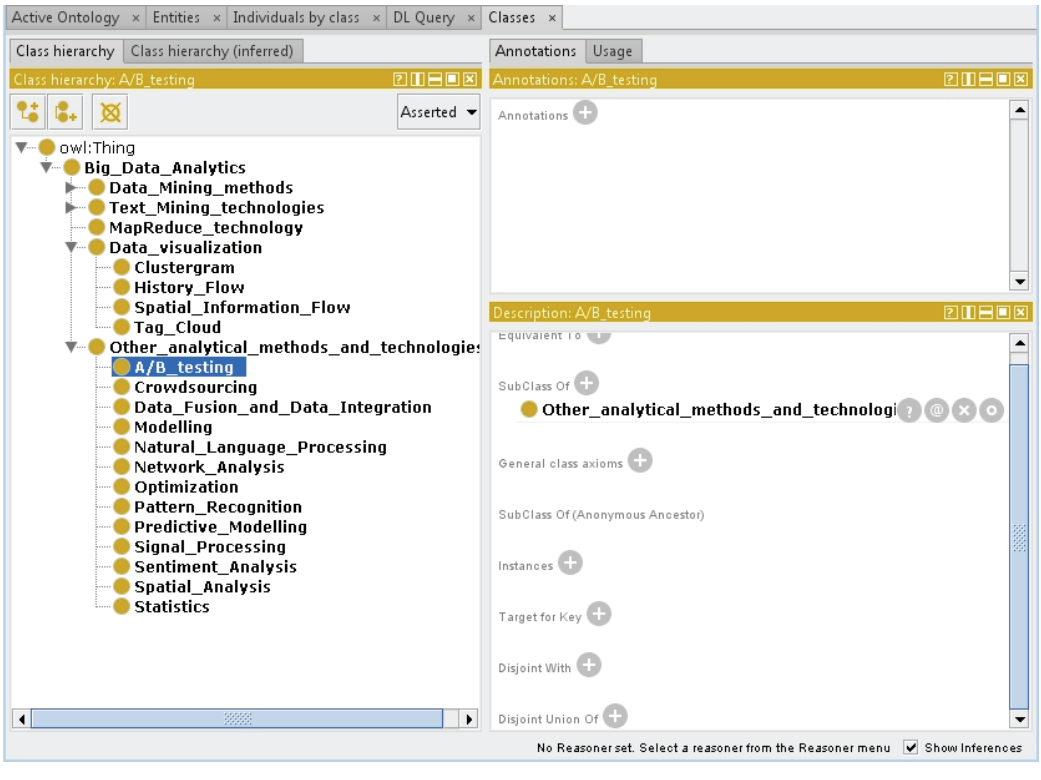

Fig. 7. Subclasses of Other technologies and research techniques class

6. Research results

The description of methods and technologies of analysis BD allows to build an ontology in accordance with the METHONTOLOGY approach [46–50], which reflects the process of iterative design. According to METHONTOLOGY methodology, the glossary of terms contains all terms (concepts and their instances, attributes, actions), important for BD analysis, and their natural language descriptions. The glossary of the terms of the ontology of the BD analysis contains the above dates, which can be divided into three groups:

1) the structure of the task (groups of technologies of analytics, communications);

ISSN 2226-3780

2) the data filling the problem (the methods used for each group);

3) results of calculations (recommendations on the BD use to improve the effectiveness of decision-making) [50–55].

The ontology of BD analysis developed by Protege-OWL is shown in Fig. 8. Each BD has a set of parameters and criteria that determine the methods and technologies for analyzing K_U. The very purpose of BD, its structure and content determine the methods and technologies of further analysis of the analysis.

Due to the developed KB ontology of BD analysis with Protege 3.4.7 and the set of RABD rules built in them, it is possible to shorten the process of selecting the methodologies and technologies for further analysis and to facilitate the automation of the analysis process of the selected BD. So, $K_U \cup K_{BD}$ at $K_U \subset K_{BD}$ will allow to generate new set K_U for A_{BD}: determination:

$$K_U = K'_{Data\ Mining} \cup K'_{MachineLearning} \cup K'_{Visualization} \cup$$
$$\cup K'_{Text\ Mining} \cup K'_{MapReduce} \cup K'_{other},$$

where

$$K'_{Data\ Mining} \subseteq K_{Data\ Mining}, \quad K'_{MachineLearning} \subseteq K_{MachineLearning},$$
$$K'_{Visualization} \subseteq K_{Visualization}, \quad K'_{Text\ Mining} \subseteq K_{Text\ Mining},$$
$$K'_{MapReduce} \subseteq K_{MapReduce}, \quad K'_{other} \subseteq K_{other}.$$

Then let's obtain a new value:

$$A'_{BD} = \langle M'_{Data\ Mining}, M'_{Machine\ Learning}, M'_{Visualization},$$
$$T'_{Text\ Mining}, T'_{MapReduce}, T'_{other}, K_U\rangle.$$

For example, for BD analysis of social networks [55], the criteria and parameters are the composition of user profiles (texts, hypertexts, age, attitude, gender, country, number of friends, posts, activity, etc.). Also, the criterion of analysis is interaction with other users of this social network and outside it. And users are not only specific individuals, but there may be information resources and agents. Applying the developed KB ontology (Fig. 9) for the BD of the social network:

$$M'_{Data\ Mining} = fdm(Tk'_{Data\ Mining}, Md'_{Data\ Mining}, K'_{Data\ Mining}),$$

where

$$Tk'_{Data\ Mining} = \langle T_{Classification}, T_{Clustering}, T_{Sequence},$$
$$T_{LinkAnalysis}, T_{Graph\ Mining}, T_{Summarization}\rangle,$$

$$MT'_{Data\ Mining} = \langle M_{Bayesian\ Networks}, M_{Tree}, M_{Neural\ Networks}\rangle,$$

$$Md'_{Data\ Mining} = \langle MD_{St}, MD_{Cb}\rangle,$$

$$MD'_{St} = \langle MS_3, MS_4\rangle, \ MD_{Cb} = \langle MC_3, MC_4, MC_6\rangle,$$

$$M'_{Machine\ Learning} = fml(ML_5, K'_{Machine\ Learning}),$$

$$M'_{Visualization} = fmv(MV_1, MV_2, MV_3, MV_4, K'_{Visualization}),$$

$$T'_{Text\ Mining} = fmt(MT_1, MT_2, MT_3, MT_6, MT_7,$$
$$MT_8, MT_9, MT_{10}, MT_{11}, K'_{Text\ Mining}),$$

$$T'_{other} = fmo(MO_1, MO_2, MO_3, MO_4, MO_9, MO_{10},$$
$$MO_{12}, MO_{13}, K'_{other}).$$

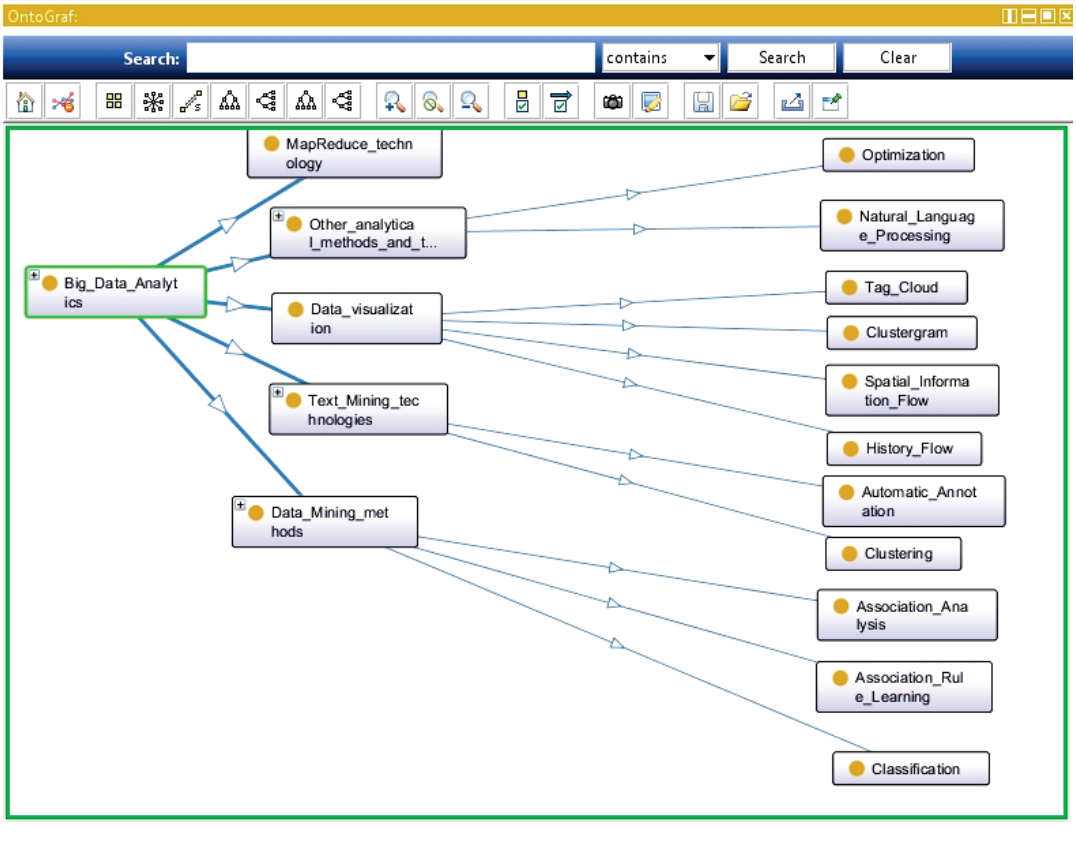

Fig. 8. The structure of the ontology for Big Data analysis as a graph

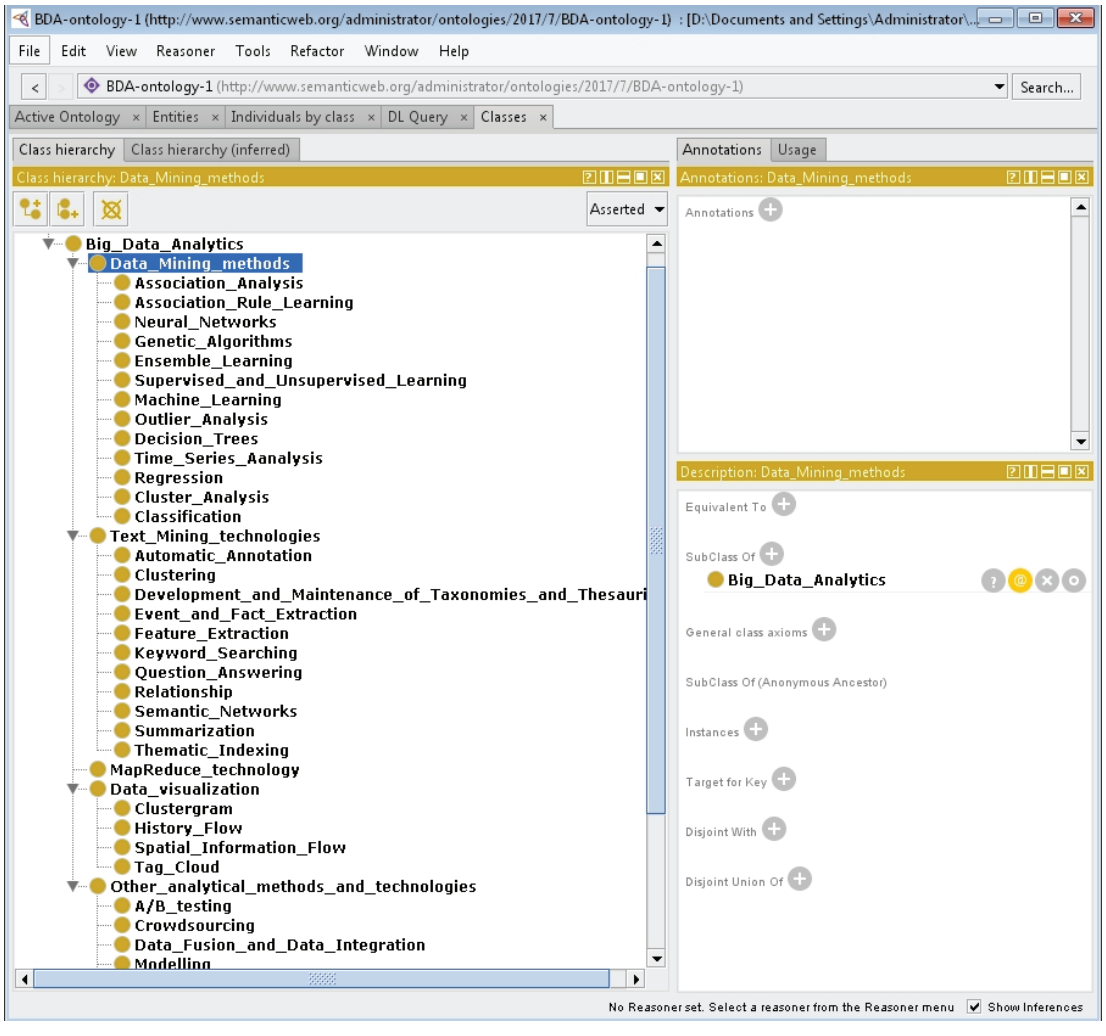

Fig. 9. Hierarchy of ontology classes for Big Data analysis

The optimal definition of all criteria and parameters of the social network analysis will allow to effectively apply the results of the analysis to identify the areas of relationships between people in many areas and in commercial activities. Nodes are network users, while links are relationships between them.

Analysis of social networks is used to solve problems of this type:

– how people from different populations form connections with outsiders;
– to find the significance or degree of influence of a particular individual in the group;
– to find the minimum number of direct links required to connect two people;
– to understand the social structure of the client base. It is possible to trace the popularization of the subject/product depending on the age, article, country of residence, status and the level of education of many users of a particular social network.

7. SWOT analysis of research results

Strengths. The ontological KB as a fundamental classifier for choosing the optimal algorithm for BD analysis of its corresponding structure and software. The classifier allows defining a set of Big Data RABD analysis rules for its use for processing and analyzing a particular BD based on its parameters and criteria.

Weaknesses. The impossibility of forming a set of Big Data RABD analysis rules in the absence of several criteria and parameters of a particular BD. An inaccurate definition of the criterion/parameter of a particular BD will lead to the formation of an inefficient BD analysis algorithm and will increase the complexity of the calculation.

Opportunities. Opportunities for further research will be to explore methods, models and tools for improving the ontology of BD analytics and effectively supporting the development of structural elements of the BD management decision support model.

Threats. Negative impact of external factors of floating of criteria and parameters set of BD analysis on the object of investigation. The absence in the world of analogues of this object of research and large-scale experiments carried out on the basis of the developed model does not give clear directions for further research.

8. Conclusions

1. The peculiarities of classification of methods and technologies of Big Data analytics are studied taking into account the definition and peculiarities of the application of the corresponding IT. The peculiarities of application

ISSN 2226-3780

of Data Mining methods, Technologies Tech Mining, Map-Reduce technology, data visualization, other technologies and BD analysis techniques allow building ontology in accordance with the METHONTOLOGY approach. It reflects the process of iterative design and allows to build a glossary of terms that are important for BD analysis and their natural language descriptions. A glossary of ontology terms for BD analysis is developed. It contains necessary terms such as task structures, task data and calculation results. The fuller the glossary, the more effective the result is in the form of an analysis algorithm BD.

2. A formal BD analysis model has been developed. At the input of the system are methods and IT analysis of BD. At the output of the system is the ontological model of BD analysis rules.

3. Ontological KB of BD analysis is developed. The taxonomy of ontology defines the BD analysis methodology. The optimal definition of the set of relations between these concepts and the set of BD analysis rules formalized with the descriptive DL logic allows efficient processing of BD.

4. Rules for Big Data RABD analysis are built. Each BD has a set of parameters and criteria that determine the methods and technologies of analysis. The very purpose of BD, its structure and content determine the methods and technologies of further analysis. Thanks to the developed KB ontology of BD analysis with Protege 3.4.7 and the set of RABD rules built in them, the process of selecting the methodologies and technologies for further analysis is shortened and the analysis of the selected BD is automated.

Acknowledgements

The work is carried out within the framework of joint research of the Department of Information Systems and Networks (ISN) of the Lviv Polytechnic National University (Ukraine) on the topic «Research, development and implementation of intelligent distributed information technologies and systems based on database resources, data warehouses, data spaces and knowledge to accelerate the formation of a modern information society». Scientific researches are also carried out within the framework of the initiative research of the Department of ICM of the Lviv Polytechnic National University on the topic «Development of intellectual distributed systems based on the ontological approach for the integration of information resources».

References

1. Mayer-Schonberger V., Cukier K. Big Data: A Revolution That Will Transform How We Live, Work, and Think. John Murray Publishers, 2013. 256 p.
2. Fekete J. D. Big Data Visual Analytics. 2016. URL: http://www.aviz.fr/wiki/uploads/TeachingVA2016/Lectur-BigDataVA.pdf (Last accessed: 18.09.2017).
3. Raghupathi W., Raghupathi V. Big data analytics in healthcare: promise and potential // Health Information Science and Systems. 2014. Vol. 2, No. 1. doi:10.1186/2047-2501-2-3
4. Hong S. H., Ma K. L., Koyamada K. Big Data Visual Analytics. NII Shonan Meeting Report No. 2015-147. Tokyo, 2017. URL: https://pdfs.semanticscholar.org/45ec/4934ee034a5839f4e657089ac865f0baa8ff.pdf (Last accessed: 18.09.2017).
5. MAD Skills: New Analysis Practices for Big Data / Cohen J. et al. // Proceedings of the VLDB Endowment. 2009. Vol. 2, No. 2. P. 1481–1492. doi:10.14778/1687553.1687576
6. History and evolution of big data analytics. URL: https://www.sas.com/en_us/insights/analytics/big-data-analytics.html (Last accessed: 18.09.2017).
7. Mitchell R. L. 8 big trends in big data analytics. URL: http://www.computerworld.com/article/2690856/big-data/8-big-trends-in-big-data-analytics.html (Last accessed: 18.09.2017).
8. Big Data. URL: http://tadviser.ru/a/125096 (Last accessed: 18.09.2017).
9. Inmon W. H. Big Data – getting it right: A checklist to evaluate your environment. Forest Rim Technology LLC. 2014. URL: http://dssresources.com/papers/features/inmon/inmon01162014.htm (Last accessed: 18.09.2017).
10. Analysis of data and processes / Barsegyan A. A. et al. Saint Petersburg: BHV-Petersburg, 2009. 512 p.
11. Paklin N. B., Oreshkov V. I. Business analysis: from data to knowledge. Saint Petersburg: Piter, 2009. 624 p.
12. Duke V., Samoylenko A. Data Mining: training course. Saint Petersburg: Piter, 2001. 368 p.
13. Manyika J. Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute, 2011. 156 p.
14. Zhuravlev J. I., Ryazanov V. V., Senko O. V. Recognition. Mathematical methods. Software system. Practical applications. Moscow: Phasis, 2006. 176 p.
15. Zinovev A. Y. Visualization of multidimensional data. Krasnoyarsk: Publisher Krasnoyarsk State Technical University, 2000. 180 p.
16. Chubukova I. A. Data Mining: A Tutorial. Moscow: Internet University of Information Technologies, BINOM, 2006. 382 p.
17. Sitnik V. F., Krasnyuk M. T. Data Mining. Kyiv: KNEU, 2007. 376 p.
18. Witten I. H., Frank E., Hall M. A. Data Mining: Practical Machine Learning Tools and Techniques. Burlington: Morgan Kaufmann, 2011. 664 p. doi:10.1016/c2009-0-19715-5
19. Marr B. Big Data: Using SMART Big Data, Analytics and Metrics to Make Better Decisions and Improve Performance. John Wiley & Sons Ltd, 2015. 256 p.
20. Einav L., Levin J. The Data Revolution and Economic Analysis. 2014. URL: http://www.nber.org/chapters/c12942.pdf (Last accessed: 18.09.2017).
21. Vanyashin A., Klimentov A., Korenkov V. PANDA follows the large data // Supercomputers. 2013. Vol. 3, No. 11. P. 56–61.
22. Serov D. Analytics of «big data» – new perspectives. URL: http://www.storagenews.ru/49/EMC_BigData_49.pdf (Last accessed: 18.09.2017).
23. Links that speak: the global language network and its association with global fame / Ronen S. et al. // Proceedings of the National Academy of Sciences. 2014. Vol. 111, No. 52. P. 5616–5622. doi:10.1073/pnas.1410931111
24. Aflalo Y., Kimmel R. Spectral multidimensional scaling // Proceedings of the National Academy of Sciences. 2013. Vol. 110, No. 45. P. 18052–18057. doi:10.1073/pnas.1308708110
25. Gadepally V., Kepner J. Big data dimensional analysis // 2014 IEEE High Performance Extreme Computing Conference (HPEC). 2014. doi:10.1109/hpec.2014.7040944
26. Analyzing Big Data with Dynamic Quantum Clustering / Weinstein M. et al. URL: https://arxiv.org/ftp/arxiv/papers/1310/1310.2700.pdf (Last accessed: 18.09.2017).
27. Paklin N. B., Oreshkov V. I. Business Intelligence: from data to knowledge. Saint Petersburg: Piter, 2013. 702 p.
28. Zelazny D. Speak in the language of diagrams: manual on visual communications for managers. Moscow: Institute for Comprehensive Strategic Studies, 2004. 220 p.
29. Roem D. The practice of visual thinking. An original method for solving complex problems. Moscow: Mann, Ivanov and Ferber, 2014. 396 p.
30. Russom P. Big data analytics. 2011. URL: https://vivomente.com/wp-content/uploads/2016/04/big-data-analytics-white-paper.pdf (Last accessed: 18.09.2017).
31. Yau N. The art of visualization in business. How to present complex information with simple images. Moscow: Mann, Ivanov and Ferber, 2013. 352 p.
32. Iliinsky N., Steele J. Designing Data Visualizations. Sebastopol: O'Reilly, 2011. 110 p.
33. Krum R. Cool infographics: effective communication with data visualization and design. Indianapolis: Wiley, 2014. 348 p.
34. Tukey J. Analysis of Observation Results: Exploratory Analysis. Moscow: Mir, 1981. 693 p.

35. Alper C., Brown K., Wagner G. R. New Software for Visualizing the Past, Present and Future. 2006. URL: http://dssresources.com/papers/features/alperbrown&wagner/alperbrown&wagner09212006.html (Last accessed: 18.09.2017).
36. Analysis of data and processes / Barsegyan A. A. et al. Saint Petersburg: BHV-Petersburg, 2009. 512 p.
37. Text Mining. URL: http://statsoft.ru/home/textbook/modules/sttextmin.html#index (Last accessed: 18.09.2017).
38. Lande D., Berezin B., Pavlenko O. Postroenie modeli informatsionnogo servisa na baze natsional'nogo segmenta Internet // Informatsionnye tehnologii i bezopasnost'. Materialy XVI Mezhdunarodnoi nauchno-prakticheskoi konferentsii ITB-2016. Kyiv: IPRI NAN Ukrainy, 2017. P. 48–57. URL: http://dwl.kiev.ua/art/itb2016/i4/i4.pdf (Last accessed: 18.09.2017).
39. Data Analysis Technologies. Data Mining, Visual Mining, Text Mining, OLAP / Barsegyan A. A. et al. Saint Petersburg: BHV-Petersburg, 2007. 384 p.
40. Linyuchev P. Text Mining: modern technologies on information mines // PC Week/RE. 2007. Vol. 6 (564). URL: https://www.pcweek.ru/idea/article/detail.php?ID=82081 (Last accessed: 18.09.2017).
41. Pleskach V. L., Zatonatskaya T. G. Information systems and technologies at enterprises. Kyiv: Znannya, 2011. 718 p.
42. MapReduce and Parallel DBMSs: Friends or Foes? / Stonebraker M. et al. // Communications of the ACM. 2010. Vol. 53, No. 1. P. 64. doi:10.1145/1629175.1629197
43. Berezin A. Map-Reduce on the example of MongoDB. 2013. URL: https://habrahabr.ru/post/184130/ (Last accessed: 18.09.2017).
44. Lebedenko E. Google MapReduce technology: divide and conquer. Kompiuterra, 2013. URL: http://www.computerra.ru/82659/mapreduce/ (Last accessed: 18.09.2017).
45. A comparison of approaches to large-scale data analysis / Pavlo A. et al. // Proceedings of the 35th SIGMOD International Conference on Management of Data – SIGMOD '09. 2009. doi:10.1145/1559845.1559865
46. Big Data from A to Ya. Part 1: Principles of working with large data, the MapReduce paradigm. 2015. URL: https://habrahabr.ru/company/dca/blog/267361/ (Last accessed: 18.09.2017).
47. Big Data from A to Ya. Part 3: Methods and strategies for developing MapReduce applications. 2015. URL: https://habrahabr.ru/company/dca/blog/270453/ (Last accessed: 18.09.2017).
48. Gavrilova T. A., Khoroshevsky V. F. Intelligent Systems Knowledge Base. Saint Petersburg: Piter, 2000. 384 p.
49. Classification Methods of Text Documents Using Ontology Based Approach / Lytvyn V. et al. // Advances in Intelligent Systems and Computing. Springer, 2016. P. 229–240. doi:10.1007/978-3-319-45991-2_15
50. Bisikalo O. V., Vysotska V. A. Identifying keywords on the basis of content monitoring method in Ukrainian texts // Radio Electronics, Computer Science, Control. 2016. Vol. 1, No. 36. P. 74–83. doi:10.15588/1607-3274-2016-1-9
51. Bisikalo O. V., Vysotska V. A. Sentence syntactic analysis application to keywords identification Ukrainian texts // Radio Electronics, Computer Science, Control. 2016. Vol. 3, No. 38. P. 54–65. doi:10.15588/1607-3274-2016-3-7
52. Lytvyn V., Bobyk I., Vysotska V. Application of algorithmic algebra system for grammatical analysis of symbolic computation expressions of propositional logic // Radio Electronics, Computer Science, Control. 2016. Vol. 4, No. 39. P. 54–67. doi:10.15588/1607-3274-2016-4-10
53. Alieksieieva K., Berko A., Vysotska V. Technology of commercial web-resource management based on fuzzy logic // Radio Electronics, Computer Science, Control. 2015. Vol. 3, No. 34. P. 71–79. doi:10.15588/1607-3274-2015-3-9
54. Matches prognostication features and perspectives in cybersport / Korobchynskyi M. et al. // Radio Electronics, Computer Science, Control. 2017. Vol. 3, No. 42. P. 95–105. doi:10.15588/1607-3274-2017-3-11
55. Wolfram S. Data Science of the Facebook World. 2013. URL: http://blog.wolfram.com/2013/04/24/data-science-of-the-facebook-world/ (Last accessed: 18.09.2017).

ОНТОЛОГИЯ АНАЛИЗА BIG DATA

Исследованы процессы анализа Big Data. Используя разработанную формальную модель и проведенный критический анализ методов и технологий анализа Big Data, построена онтология анализа Big Data. Исследованы методы, модели и инструменты для усовершенствования онтологии аналитики Big Data и эффективной поддержки разработки структурных элементов модели системы поддержки принятия решений по управлению Big Data.

Ключевые слова: онтология анализа Big Data, данные визуализации, интеллектуальный анализ данных, Text Mining, MapReduce.

Lytvyn Vasyl, *Doctor of Technical Sciences, Professor, Department of Information Systems and Networks, Lviv Polytechnic National University, Ukraine, e-mail: yevhen.v.burov@lpnu.ua, ORCID: http://orcid.org/0000-0002-9676-0180*

Vysotska Victoria, *PhD, Associate Professor, Department of Information Systems and Networks, Lviv Polytechnic National University, Ukraine, e-mail: victoria.a.vysotska@lpnu.ua, ORCID: http://orcid.org/0000-0001-6417-3689*

Veres Oleh, *PhD, Associate Professor, Department of Information Systems and Networks, Lviv Polytechnic National University, Ukraine, e-mail: Oleh.M.Veres@lpnu.ua, ORCID: http://orcid.org/0000-0001-9149-4752*

Brodyak Oksana, *PhD, Associate Professor, Department of Mathematics, Lviv Polytechnic National University, Ukraine, ORCID: http://orcid.org/0000-0002-9886-3589*

Oryshchyn Oksana, *PhD, Associate Professor, Department of Mathematics, Lviv Polytechnic National University, Ukraine, e-mail: oksana.orushchun@gmail.com, ORCID: http://orcid.org/0000-0002-8965-1891*