Big data sets in construction

Pavel Kagan[1,*]

[1]Moscow State University of Civil Engineering, Yaroslavskoe shosse, 26, Moscow, 129337, Russia

Abstract. The paper studies the processing of large information data arrays (Big Data) in construction. The issues of the applicability of the big data concept (Big Data) at various stages of the life cycle of buildings and structures are considered. Methods for data conversion for their further processing are proposed. The methods used in the analysis of "big data" allow working with unstructured data sets (Data Mining). An approach is considered, in which the analysis of arbitrary data can be reduced to text analysis, similar to the analysis of ordinary text messages. At the moment, it is important and interesting to isolate non-obvious links present in the analysed data. The advantage of using big data is that it is not necessary to advance hypotheses for testing. Hypotheses appear during data analysis. Dependence analysis is a basic approach when working with big data. The concept of an automatic big data analysis system is proposed. For data mining, text analysis algorithms should be used, and discriminant functions should be used for the main problem to be solved (data classification).

1 Introduction

In recent years, firms, companies, enterprises, public services, etc. due to its activities, a considerable amount of data, sometimes heterogeneous and unstructured, has been accumulated, on different types of media, with different storage methods, etc. At the same time, the challenges of time urgently require that the accumulated data benefit, for example, firms - profit, public services, police - operational and archival information of various nature, the banking sector - data on the possibility of issuing loans, etc.

Nowadays, the approaches associated with processing, storing and analyzing large amounts of data have become quite widespread. Such a direction, connected with the change of data processing and analysis technology, with the development of distributed data storage and processing systems, with the departure from traditional bases and banks, was given the general name Big Data [1, 2].

Traditional ways of processing information presented in the form of relational databases are not able to work with unstructured data, such as free text or data coming from analog sensors. Therefore, another important concept in this area was the term Data Mining ("intelligent" or "in-depth" data analysis), that is, a set of methods for detecting previously unknown, non-trivial, but useful and accessible knowledge in large data arrays. for decision making in various areas of human activity.

* Corresponding author: pavel501@rambler.ru

The term “DataMining”, which is often translated as "mining" or "excavating data", describes a system for finding patterns in data and, possibly, predicting trends in their occurrence. The definition used most frequently is one of the founders of this direction G. Piatetsky-Shapiro (GTE Labs): "DataMining is the process of finding in raw data of previously unknown non-trivial practically useful and accessible interpretations of knowledge necessary for making decisions in various areas of human activities". Similar problems have arisen in connection with the huge amount of information of various types accumulated in databases. Often one can come across a synonym for DataMining - “knowledge discovery in databases”, emphasizing the need not only to extract, but also to analyze the information contained in the databases. It should be noted that the term "raw", which characterizes the data in the definition, reflects not the fact that the data are not structured, but precisely that no useful information is extracted from them. In addition, since the process of extracting and analyzing information is quite laborious, very often such work requires the use of large computational powers.

To unify the algorithms used to work with big data, the data can be represented as words of the language and text analysis algorithms can be used. By transforming quantitative and ordinal variables into nominal variables, you can create a fully automated system for using data from different spheres of human activity and data that are variables of various types.

In principle, the scope of application of DataMining methods is not limited to the above areas, the methods can be applied in any area where data is available. Related events (association), sequence of related events, classification (assignment of an object to a certain class), clustering (identifying homogeneous groups of data) and forecasting are standard types of patterns, detected using DataMining methods.

Possible areas of application include insurance and banking (development of tariff plans and fraud detection), medicine (to identify patterns in diagnosis and prescription of treatment), demography, analysis of the consumer basket and building predictive consumption patterns, as well as many others.

Work with big data and Data Mining were developed primarily in areas such as:

- analysis of various information on the Internet;
- sociology - when conducting various surveys, surveys;
- statistical studies;
- the medicine;
- analysis of the securities market;
- analysis of the real estate market;
- analytical activities in the banking sector;
- education;
- retail;
- tourism;
- and etc.

Big data analysis allows stakeholders and investors to identify new trends and business opportunities. Big data and Data Mining provide a broader understanding of what drives investment in the market and allows you to plan your corporate strategy.

There are various algorithms that are used in DataMining, for example, decision trees that create a hierarchical structure of the rules "if ... then ...", limited search algorithms, artificial neural networks, genetic algorithms, fuzzy logic algorithms, etc. [15]

For example, a genetic algorithm (this is a simple model of evolution in nature, implemented as an algorithm) is mainly used to solve combinatorial and optimization problems [16]. The algorithm consists in the following: let some objective function be given, which in general depends on several variables, and it is required to find the values of the variables for which the value of the function is maximal. The algorithm uses both an analogue of the mechanism of genetic inheritance and an analogue of natural selection. This uses biological

terminology. We are dealing with an individual (individual). An individual is some solution to a problem. An individual is considered to be more adapted, the better the corresponding solution (the greater the value of the objective function this solution gives). Then, choosing the most adapted individual in the current generation, you can get a not absolutely accurate, but close to optimal answer. Individuals are endowed with chromosomes.

The genetic algorithm mimics the evolution of a population of individuals as a cyclic process of selecting and crossing chromosomes (a vector containing a set of values) and changing their generations, which continues until a given number of generations change or some other stopping criterion is fulfilled. During the life of a population, random crosses occur (crossover operation, in which two chromosomes exchange their parts) and mutations (random change of one or several positions in the chromosome), as a result of which new chromosomes appear.

Artificial neural networks can also be successfully applied to DataMining tasks. Artificial neural networks are simplified models of biological neural networks of the brain of living beings, and in these models with a large number of parallel working fairly simple calculators “synapse (multiplier) - adder - threshold element” with a high speed solves quite complex tasks. These tasks include classification, clustering, the search for patterns, associations, etc. In particular, self-organizing maps (Kohonen networks) can improve the understanding of the data structure, which will help to more effectively carry out exploratory data analysis, detect new phenomena, etc. Other examples of tasks solved by artificial neural networks are the prediction of sales of products, the provision of services, indicators of the exchange market, etc.

Decision trees are one of the most visible and powerful methods of data analysis in terms of studying the relationship of one dependent and several independent variables (predictors). This method allows you to set the specified relationship not using a predictive equation (as opposed to regression analysis), but using hierarchical data segmentation, which ultimately form a tree structure. Decision trees are based on machine learning, decision trees are based on decision rules of the form "if ... then ...".

The main types of decision trees are as follows:

The classification tree is used to assign objects to one of the previously known classes and occurs when the probability of a categorical value is predicted (that is, a variable, each value of which indicates that the object belongs to a certain group (category)) dependent variable according to the corresponding predictor values.

A regression tree occurs when it is necessary to predict the average value of a quantitative dependent variable for the corresponding predictor values.

To build a tree, the entire training data set is taken, divided into two or more nodes so that the observations that fall into different nodes are as different as possible in the dependent variable. The partitioning rules that maximize these differences are the values of independent variables.

To build the underlying nodes of the tree (we will talk about descending trees that are used more often) it is necessary at each node of this level to find such a criterion for splitting the set of statistical data associated with this node so that the resulting subsets consist of elements of the same class. The quality of the partition is evaluated using statistical criteria. The partitioning process continues until the appearance of terminal nodes, that is, nodes that cannot be split further. Analysis of the terminal nodes of the tree allows you to find the optimal answer.

2 Materials and Methods

Big data in the construction and operation of real estate.

Due to modern innovative technological changes (hardware and software) for processing big data, it becomes possible to use similar methods in other areas, such as construction and housing and public utilities (for example, [3, 4]). Big data analysis can reveal opportunities for improving various aspects of design, construction and operation. A variety of input data allows you to increase the level of reliability of status reports and forecasts, issue warnings when acceptable indicators are exceeded. The wide dissemination of models built on the basis of large amounts of evidence eliminates the limitations caused by not entirely accurate hypotheses and assumptions in the models.

Consider a set of areas of theoretical and practical research that can and will be developed in this area of activity.

Due to the widespread concept of "life cycle" [5] of building objects and the formation of BIM-models (Building Information Model) [6,7,8] of such objects, further presentation can be constructed in accordance with the main stages of "life" such objects.

1. Conceptual stage

Data analysis (big data and Data Mining) makes it possible to evaluate the design features of the future building, the choice of building materials used, the location of the object (climatic and geological features, traffic and business activity), the expected profit from the project, the environmental situation, the project's image in social networks and more. Analysis of the history of the implementation of similar projects reveals patterns and likelihood of construction risks.

Possible directions for the development of data mining to improve the efficiency of solutions at the conceptual design and planning stage (creating the "concept" of a future project / object) related to big data:

- Forecasting the development of the future "living environment" of a person within a specific territory;
- Forecasting requirements for real estate (including the development of the social environment, changing people's needs, the development of territories [9,10], changing environmental and natural factors, etc.);
- forecasting the future "demand" of real estate;
- and etc.

2. Design stage

Possible directions for the development of data mining to improve the efficiency of solutions at the design stage associated with big data:

- formation of expert systems for analyzing design solutions;
- development of the concept of alternative design based on BIM-models using in-depth / intellectual analysis of project documentation;
- use of the in-depth / intellectual analysis system for finding "hidden knowledge" (previously unknown, nontrivial or practically useful) for the synthesis and evaluation of the effectiveness of design solutions in the design process.

Related tasks (problems):

- determination of the requirements for the initial information from BIM-models (at the design stage), which can be used for data analysis;
- detection of "collisions" (contradictions) in BIM-models;
- analysis of project risks;
- and etc.

3. Stage of construction

Data analysis has always been vital to ensure efficient construction production, and now there are tools that enable managers to evaluate and use resources more efficiently.

For a long time, architects and designers created their projects only on paper and there was no exact digital model. When creating a digital model, it often turns out that "as built" is very different from the project. The risk assessment for the existing deviations from the

project that actually affect the safety of the facility becomes relevant. Modern methods of working with data and progress in the development of computing technology allow us to analyze a large number of factors for more accurate classification by risk groups. Thus, minor project inconsistencies that do not lead to significant risks do not require rework and additional work. This allows you to achieve economic effect when entering new facilities and operating existing ones.

Building materials and equipment can come from different places, including remote plants and other construction sites. The availability of this data allows the manager to track the location and composition of these resources, which allows you to make effective decisions about their use. This allows you to quickly move resources to a new location. Tracked the use of technology and the justification of its movement. Big Data and Data Mining can also help select subcontractors and suppliers with the best reputation.

At the stage of construction of construction objects (construction and installation works), the following can be used as "natural" data sources:
- data of video registration of surveillance cameras installed on the construction site;
- data obtained from sensors (sensors) for various purposes, fixing the technological parameters of the erected structural elements.

Possible directions of creating data mining systems for the synthesis and evaluation of the effectiveness of management decisions at the construction stage, associated with big data:

- data processing of video signals of surveillance cameras installed at the construction site. Collecting data directly from workstations allows you to properly organize workflow and minimize downtime. Such an approach can solve problems of labor discipline and increase productivity;
- data processing from sensors (sensors);
- analysis of performance documentation;
- analysis of data from BIM-models;
- risk assessment of deviations from projects at the construction stage;
- and etc.

Related tasks (problems):
- definition of requirements for the initial information from BIM-models (at the design stage), which can be used at the stage of construction works;
- determination of requirements for structuring and "accumulating" information of BIM-models at the stage of construction of construction objects;
- and etc.

3 Results

The main methods of solving the above tasks at different stages and stages of the "life cycle" of construction objects are the use of: artificial neural networks, genetic algorithms, decision trees, limited search algorithms, fuzzy logic algorithms, etc. [12, 13].

In this study, it is proposed to use text analysis algorithms for solving such problems. Since the main problem to be solved is the classification problem, discriminant functions are used. An example of an action scheme (algorithm) that provides processing of large volumes of information using text analysis algorithms (see Fig. 1).

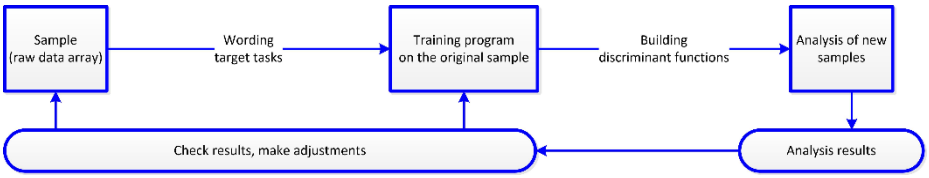

Fig. 1. Functional diagram of the procedure for analytic processing of big data using text analysis algorithms.

In practice, the calculation of discriminant functions is a heavy computational problem, requiring the use of competitive programming methods. Analysis of a new sample using Apache Spark [14] of 1000 elements takes a few seconds on a personal computer.

Thus, problems and tasks, the solution of which is associated with the analysis of big data, are widespread.

It should be noted that at the moment the leading economies of the world are paying great attention to the use of technologies associated with big data in the construction industry. Already at the feasibility stage, the use of big data provides a better understanding of costs and timing. A large number of analyzed data can significantly reduce the risk of projects. Identifying patterns, in order to process current data in real time, allows you to better control production and stick to schedules.

The article proposes to use text analysis algorithms for data mining, and to use discriminant functions for the main task to be solved (classification problem).

4 Discussion

Stage of operation

Data collection and analysis technologies strive to capture various workflows, make data available for analysis and allow you to create new business solutions. Over the past few decades, research has been conducted on the development of methods for the detection and diagnosis of faults and the application of knowledge about construction sites to improve maintenance during operation.

Data from sensors embedded in buildings, bridges, any objects and structures, allow you to control various parameters. For example, utility companies have sophisticated billing analytics systems, but they may not have a set of tools or opportunities for more in-depth analysis of how electricity is consumed by various equipment and systems in buildings or even for correlating employment, weather conditions and operating modes of various temperature control equipment. , humidity and power consumption. Energy saving control in shopping centers, offices and residential buildings can be carried out in real time.

Along with the analysis of information from technical means, information on the Internet is of interest. Analysis of text messages allows you to analyze the satisfaction of residents with the quality of services provided.

At the stage of operation of construction sites, the following can be used as “natural” data sources:

- data of video registration of surveillance cameras;
- data metering devices installed in various parts of engineering networks [11];
- data obtained from sensors (sensors) for various purposes, fixing the parameters of the operated structural elements and premises.

Possible directions of creating data mining systems for the synthesis and evaluation of the effectiveness of management decisions at the operational stage associated with big data:

- data processing of video signals of security systems;
- data processing metering devices;
- data processing from sensors (sensors);
- analysis of operating documents;
- analysis of data from BIM-models;
- risk analysis, planning and forecasting;
- and etc.

Related tasks (problems):

- determination of requirements for data and data sources (video and audio signals, information of sensors and metering devices, the number and location of these sources, etc.);
- definition of requirements for the initial information from BIM-models (at the stages of design and construction), which can be used at the operation stage;
- determination of requirements for structuring and "accumulating" information of BIM-models at the operational stage;
- and etc.

References

1. D. Forman, *Many numbers: analyze big data using Excel* (Al'pina Pablisher Publ., Moscow 2016)
2. A. Prokopets *Competitive programming on Scala.* (DMK-Press Publ., Moscow 2018)
3. O. Shestakova, L. A. Kychkin, Vestnik MGSU **10**, 1191–1201, (2017)
4. A. Konikov, G. Konikov, Promyshlennoe i grazhdanskoe stroitel'stvo **10**, 78–82 (2017)
5. A. Ginzburg, *Building Life Cycle Information Modelling.* **9**, 61–65 (2016)
6. A. Volkov, L. Sukneva, *BIM-Technology in Tasks of the Designing Complex Systems of Alternative Energy Supply*. Procedia Engineering **23**, 377-380 (2014)
7. P. Kagan, S. Muminova, *Bim training course in construction university* In: Proceedings of the 11th International Conference on Construction Applications of Virtual Reality, pp. 72-77. Bauhaus-Universität, Weimar (2011).
8. A. Ginzburg, L. Shilova, A. Adamtsevich, L. Shilov, Journal of Applied Engineering Science **14(4)**, 457-460 (2016)
9. P. Kagan, Vestnik KIGIT **9**, 12-3 (2012)
10. B. Pavel, V. Kulikov, *Information modeling of urban planning development*, Applied Mechanics and Materials 409-410, 951-954 (2013)
11. P. Kagan: Procedia Engineering **25**, 261-265 (2016)
12. P. Kagan, R. Polyakov, Nauchnoe obozrenie **10**, 15–19 (2017)
13. V. Ignatov, E. Ignatova, Vestnik MGSU **4**, 332–335, (2009)
14. U. Lezerson, S. Riza, *Spark for professionals*. Modern patterns of big data processing. (Izdatel'skiy dom «Piter» Publ., St. Petersburg, 2017)
15. A. Sviridov, V. Simonov, S. Alkadarsky, and etc., *Fuzzy and neuro-fuzzy systems and technologies*. (Izd. RSSU, Moscow, 2011)
16. N. Kuzyurin, S. Martishin, M. Khrapchenko, *Genetic Algorithms in the Search Problem for Commonly Found Combinations*. Proceedings of the Institute of System Programming of the Russian Academy of Sciences **6**, 109-126 (2004)
17. A. Gruzdev, *Predictive modeling in IBM SPSS Statistics and R. Method of decision trees*. (DMK-Press Publ., Moscow, 2016)