ANALYSIS OF THE USE OF MINING METHODS AND MODELS IN DATA PROCESSING

Daminova Barno Esanovna
Associate Professor at Karshi state university,
Orchid Number: 0009-0001-4211-6082
daminova.barno75@gmail.com

Annotation. *The article discusses the development of information technologies, data collection and storage systems, which are used to solve problems of analyzing large volumes of information.*

Key words: *information and communication technologies, process, analysis, intellectualization, data analysis.*

In the process of development of information technologies, as well as systems for collecting and storing data, the problem of analyzing large volumes of information is becoming increasingly acute. Another equally important task is the task of visual and compact presentation of data. These problems are solved within the framework of an interdisciplinary field of knowledge - data mining.

Data mining refers to processing information and identifying trends in it that helps make decisions. There are many different methods for data mining, query processing modeling, and information gathering. Data mining (many are familiar with this term by its English name data mining) is one of the most pressing topics in the modern world.

The principles of data mining have been known for a long time, but with the advent of big data they have become even more widespread. When working with large data sets, relatively simple and straightforward statistics are no longer enough. With about a million detailed records, it is necessary to know not only information about the location of several thousand records, but also to understand whether this category of records belongs to a certain group.

These requirements have created a complex data mining process. Solving problems requires data analysis, which builds a model to describe the information and finally creates a resulting report.

Clustering is a task at first glance very similar to the classification method. This method, also called cluster analysis, is a logical extension or, if you like, a generalized case of a classification problem where the set of classes is not known in advance.

Data mining can be performed with relatively modest database systems and simple tools, including building your own or using off-the-shelf software packages. Sophisticated data mining relies on past experience and algorithms defined using existing software and packages, with different specialized tools associated with different methods.

Today, the analysis of data obtained from the Internet, the so-called Web Mining, is becoming increasingly relevant. The main goal of Web Mining is collecting data (parsing) and then saving it in the required format. It is necessary to take into account that information on

the web is presented in the form of special formats, such as markup language HTML, RSS, Atom, SOAP, etc. Web pages may have additional metainformation, as well as information about the structure of the document.

In Web Mining, two main directions can be distinguished: Web Content Mining and Web Usage Mining, and, accordingly, two types of tasks that are posed to Web Mining systems. Web Content Mining means the automated search of information from various sources on the Internet. The second direction is more suitable, Web Usage Mining involves the detection of patterns in the actions of website visitors, as well as the collection of statistics and its subsequent analysis.

One of the main methods for solving the problem of Web Content Mining, namely an algorithm for hierarchical clustering of data from Internet sources. Cluster analysis, the principles of which were used when constructing the algorithm, does not require a priori assumptions about the source data, does not impose restrictions on the representation of the objects under study, and allows one to analyze indicators of various types of data.

Unlike classification problems, solving clustering problems is based on comparing the objects themselves and establishing their similarity according to certain characteristics. The developed algorithm was applied to a real example of combining blogs containing semantically similar posts.

Clustering algorithms divide a collection of data into subsets, or clusters.

The goal of these algorithms is to create clusters that are internally homogeneous but clearly distinct from each other. Data or vectors of characteristics that are elements of a set within a cluster should be as similar to each other as possible, but at the same time as different as possible from the elements of another cluster.

Clustering is the most common form of unsupervised learning. The absence of a teacher means that the algorithm does not involve an expert assigning elements to classes. In clustering problems, the distribution and structure of data determine cluster membership. The main input information for the clustering algorithm is the metric. When clustering texts, Euclidean distance is often used.

Changing the metric can often have a large impact on the clustering results. Thus, the metric is a very important tool with which you can change the clustering results. It is worth noting that flat clustering generates a set of clusters that do not have obvious relationships. Hierarchical clustering creates a hierarchy of clusters.

Flat clustering, which produces a set of clusters that have no explicit relationships, is efficient and simple, but the result is a simple, unstructured set of clusters using the number of clusters as an input parameter. Hierarchical clustering creates a hierarchy, that is, a structured set that is more informative than an unstructured set of clusters. Hierarchical clustering does not require you to specify the number of clusters you want in advance, but these benefits of hierarchical clustering come at the cost of lower performance. The complexity of the most common hierarchical clustering algorithms is at least quadratic with respect to the number of input data.

Traditionally, two main types of hierarchical clustering algorithms are used: top-down or bottom-up. Bottom-up algorithms initially consider each input element as a separate cluster,

and then sequentially combine (agglomerate) the clusters. Depending on the parameters of the algorithm, this merging will occur until either one cluster is formed containing all the elements, or a predetermined number of clusters is reached. Top-down clustering initially considers a single cluster of all elements and then recursively splits it until individual objects are obtained.

Predictive methods use the values of one variable to predict unknown (target) values of the variables. When combined with other data mining techniques, forecasting involves trend analysis, classification, model matching, and relationships.

Sequential models are used to analyze long-term data, a useful method for identifying trends or regular occurrences of similar events. The decision tree, associated with most other methods, is used as part of the selection criteria and to support the selection of specific data within an overall framework. A decision tree starts with a simple question that has two answers (but possibly more).

Each answer leads to the next question, helping you classify and identify data or make predictions.

Decision trees are most often used with information classification systems and forecasting systems, where various forecasts can be based on past historical experience, which helps build the decision tree structure and obtain the result. The combination method is used quite rarely. It is similar to classification and clustering methods.

With all basic methods, it often makes sense to record and later study the information obtained. For some methods this is quite obvious. For example, when building sequential models and training, historical data from different sources and instances of information is analyzed for forecasting purposes.

In other cases, this process may be more pronounced. Decision trees are rarely built once and never forgotten. As new event information and data points are identified, additional branches or even entirely new trees may need to be built.

Some of these processes can be automated. For example, building a predictive model to detect credit card fraud involves determining the probabilities that can be used for the current transaction and then updating the model as new (confirmed) transactions are added. This information is then recorded so that a decision can be made faster next time.

Data mining relies on building a suitable model and structure that can be used to process the identification and creation of the required information. Regardless of the form and structure of the data source, information is structured and organized according to a format that allows data mining to be performed with the most efficient model.

Analytical variables for data obtained from many different sources can be compiled into a single specific structure (for example, create a class of buyers of certain levels and ages, or a class of errors of a certain type).

Depending on the data source, it is important to choose the right way to construct and transform this information, whatever the final data analysis method. This step also leads to a more complex process of identifying the collection of simplification or expansion of information according to the input data.

Data source location and database influence how information will be processed and combined.

Various Data Mining methods are characterized by certain properties that can be decisive when choosing a data analysis method. Methods can be compared with each other by assessing the characteristics of their properties.

The main properties and characteristics of Data Mining methods: accuracy, scalability, interpretability, verifiability, complexity, flexibility, speed and popularity.

Scalability is a property of a computing system that provides predictable growth in system characteristics, such as responsiveness, overall performance, etc., when computing resources are added to it.

Data mining is not just about running some complex queries on data stored in a database. Regardless of whether you are using a SQL database based on documents, simple flat files, you need to work with the data to format or restructure it. You need to determine the format of information on which your method and analysis will be based. Then, once the information is found in the desired format, various techniques can be applied (individually or in combination) independent of the required underlying data structure or data set.

Working with large data sets, as well as their processing, makes it possible to create complex generalizations of the results of data mining by groups and data comparisons. New systems and tools are now available, such as combined data storage and processing systems.

SQL databases are strictly structured and adhere to a rigid schema, making them easy to query and analyze data with a known format and structure.

REFERENCES:

1. Nurfahasdi M. et al. Optimization of Efficiency Mercury (Hg) Removal with Electrocoagulation Using Zinc (Zn) Electrode by RSM Methods //E3S Web of Conferences. – EDP Sciences, 2024. – Т. 497. – С. 01002.
2. Benzerara M. et al. Advanced Strengthening of Steel Structures: Investigating GFRP Reinforcement for Floor Beams with Trapezoidal Web Openings //E3S Web of Conferences. – EDP Sciences, 2024. – Т. 497. – С. 02013.
3. Student M. D. et al. THE ROLE OF MODERN INFORMATION AND COMMUNICATION TECHNOLOGIES IN TEACHING LESSONS IN MATHEMATICS AND COMPUTER SCIENCE //Экономика и социум. – 2024. – №. 2-2 (117). – С. 88-93.
4. Daminova B. FORMATION OF THE MANAGEMENT STRUCTURE OF EDUCATIONAL PROCESSES IN THE HIGHER EDUCATION SYSTEM //Science and innovation. – 2023. – Т. 2. – №. A6. – С. 317-325.
5. Ergash o’g’li Q. F., Jumanazarovna B. I. METHODS OF DISPLAYING MAIN MEMORY ON CACHE //Ответственный редактор. – 2020. – С. 6.
6. Даминова Б. Э. СОДЕРЖАНИЕ ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ И ТЕНДЕНЦИИ ЕГО ИЗМЕНЕНИЯ ПОД ВЛИЯНИЕМ НОВЫХ

СОЦИАЛЬНО-ЭКОНОМИЧЕСКИХ УСЛОВИЙ //Yosh mutaxassislar. – 2023. – Т. 1. – №. 8. – С. 72-77.

7. Бозорова И. Ж. МЕСТО И РОЛЬ ИНФОРМАЦИОННО-КОММУНИКАЦИОННЫХ ТЕХНОЛОГИЙ В ЭКОНОМИЧЕСКИХ ПРОЦЕССАХ //Экономика и социум. – 2024. – №. 2-1 (117). – С. 910-915.

8. Daminova B. E. GAUSS AND ITERATION METHODS FOR SOLVING A SYSTEM OF LINEAR ALGEBRAIC EQUATIONS //Экономика и социум. – 2024. – №. 2 (117)-1. – С. 235-239.

9. Bozorova I. J. METHODS OF PROCESSING AND ANALYSIS OF BIO SIGNALS IN ELECTROCARDIOGRAPHY //ПРОБЛЕМЫ СОВРЕМЕННЫХ ИНТЕГРАЦИОННЫХ ПРОЦЕССОВ И ПОИСК ИННОВАЦИОННЫХ РЕШЕНИЙ. – 2020. – С. 97-99.

10. Daminova B. E., Oripova M. O. METHODS OF USING MODERN METHODS BY TEACHERS OF MATHEMATICS AND INFORMATION TECHNOLOGIES IN THE CLASSROOM //Экономика и социум. – 2024. – №. 2 (117)-1. – С. 256-261.

11. Бозорова Н. Ж., Азизова Н. Б. Технология Продуктивного Чтения: Приёмы Предтекстовой Деятельности В Обучении Русскому Языку Как Иностранному //International Journal of Formal Education. – 2024. – Т. 3. – №. 1. – С. 222-224.

12. Кайнаров Ф. З. ИННОВАЦИОННЫЕ МЕТОДЫ ПРЕПОДАВАНИЯ ПРИКЛАДНОЙ МАТЕМАТИКИ //Экономика и социум. – 2023. – №. 1-2 (104). – С. 619-622.

13. Bozorova I. J., Sh M. F., Rustamov M. A. NEURAL NETWORKS. NEURAL NETWORKS: TYPES, PRINCIPLE OF OPERATION AND FIELDS OF APPLICATION //РОЛЬ ИННОВАЦИЙ В ТРАНСФОРМАЦИИ И УСТОЙЧИВОМ РАЗВИТИИ СОВРЕМЕННОЙ. – 2020. – Т. 130.

14. Bozorova I. J., Zoxidov J. B., Turdiyeva M. A. STORAGE OF BIOMEDICAL SIGNALS AND FORMATS OF BIOSIGNALS //СОВЕРШЕНСТВОВАНИЕ МЕТОДОЛОГИИ И ОРГАНИЗАЦИИ НАУЧНЫХ. – 2020. – Т. 116.

15. Маматмурадова М. У., Бозорова И. Ж., Кодиров Ф. Э. ПРОБЛЕМЫ СОВРЕМЕННЫХ ПРОГРАММНЫХ И КОМПЬЮТЕРНО-ИНЖЕНЕРНЫХ ТЕХНОЛОГИЙ И СОВРЕМЕННЫЕ ТЕХНОЛОГИИ СОЗДАНИЯ ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ //Инновации в технологиях и образовании. – 2019. – С. 294-297.

16. Esanovna D. B. et al. ELECTRONIC TEXTBOOK AS A BASIS FOR INNOVATIVE TEACHING //MAVZUSIDAGI XALQARO ILMIY-AMALIY ANJUMAN. – С. 660.

17. Даминова Б. Э. и др. ОБРАБОТКА ВИДЕОМАТЕРИАЛОВ ПРИ РАЗРАБОТКЕ ОБРАЗОВАТЕЛЬНЫХ РЕСУРСОВ //Экономика и социум. – 2024. – №. 2-2 (117). – С. 435-443.

18. Даминова Б. Э., Якубов М. С. Проблемы защиты от внешних и внутренних информационных угроз //Труды Северо-Кавказского филиала Московского технического университета связи и информатики. – 2013. – №. 1. – С. 306-308.

19. Daminova B. ACTIVATION OF COGNITIVE ACTIVITY AMONG STUDENTS IN TEACHING COMPUTER SCIENCE //CENTRAL ASIAN JOURNAL OF EDUCATION AND COMPUTER SCIENCES (CAJECS). – 2023. – Т. 2. – №. 1. – С. 68-71.

20. Даминова Б. Э. СРАВНИТЕЛЬНЫЙ АНАЛИЗ СОСТОЯНИЯ ОРГАНИЗАЦИИ МНОГОУРОВНЕВЫХ ОБРАЗОВАТЕЛЬНЫХ ПРОЦЕССОВ //Экономика и социум. – 2023. – №. 1-2 (104). – С. 611-614.