ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative Technologies and Scientific Solutions for Industries. 2020. No. 3 (13)

UDC 004.75 DOI: https://doi.org/10.30837/ITSSI.2020.13.076

V. FILATOV, V. SEMENETS, O. ZOLOTUKHIN

DATA MINING IN RELATIONAL SYSTEMS

The **subject** of the research is methods of relational database mining. The **purpose** of the research is to develop scientifically grounded models for supporting intelligent technologies for integrating and managing information resources of distributed computing systems. Explore the features of the operational specification of the relational data model. To develop a method for evaluating a relational data model and a procedure for constructing functional associative rules when solving problems of mining relational databases. In accordance with the set research goal, the presented article considers the following **tasks**: analysis of existing methods and technologies for data mining. Research of methods for representing intelligent models by means of relational systems. Development of technology for evaluating the relational data model for building functional association rules in the tasks of mining relational databases. Development of design tools and maintenance of applied data mining tasks; development of applied problems of data mining. **Results**: The analysis of existing methods and technologies for data mining is carried out. The features of the structural specification of a relational database, the formation of association rules for building a decision support system are investigated. Information technology has been developed, a methodology for the design of information and analytical systems, based on the relational data model, for solving practical problems of mining, practical recommendations have been developed for the use of a relational data model for building functional association rules in problems of mining relational databases, **conclusion**: the main source of knowledge for database operation can be a relational database. In this regard, the study of data properties is an urgent task in the construction of systems of association rules. On the one hand, associative rules are close to logical models, which makes it possible to organize efficient inference procedures on them, and on the other hand, they more clearly reflect knowledge than classical models. They do not have the strict limitations typical of logical calculus, which makes it possible to change the interpretation of product elements. The search for association rules is far from a trivial task, as it might seem at first glance. One of the problems is the algorithmic complexity of finding frequently occurring itemsets, since as the number of items grows, the number of potential itemsets grows exponentially.

Keywords: information system; database; relational data model; data integration; intelligent systems; extracting knowledge from data; data mining; associative patterns of data.

Introduction

Knowledge Discovery in Databases (KDD) methods have been actively researched and developed over the past 20 years. Data mining technologies that are currently used are the result of the evolution of the following directions: in-depth development, intellectualization, increasing the level of mathematical methods for data processing, further development of integrated systems – databases and decision support systems, modeling of nervous tissue of animals and humans with artificial neural networks [1].

Currently, there are many large research centers and teams engaged in the development of methods and the creation of KDD systems. Among the large companies that are intensively dealing with this issue, one can single out IBM and Microsoft. IBM has completely repurposed its largest software technology research center in Almaden to develop KDD algorithms and build working KDD systems. The result of this work is a whole family of KDD systems, both general purpose and specialized. For example, one of the specialized systems, called Advanced Scout, is used by the US Basketball Association to analyze the effectiveness of various combinations of players in teams, to analyze game situations and to develop a game strategy. This specialized system costs over a million dollars and is used by several IBA teams. Microsoft has set up KDD, located directly at the firm's headquarters in Redmond, and has brought in renowned experts who have previously worked on the subject at universities and academic research centers. This center is headed by Professor Osama Fayadd, who received one of the most honorable American awards for the advancement of science in 1996, and formerly worked in the NASA Jet Propulsion Laboratory.

These problems are also dealt with at universities. One of the oldest research groups is at Wichita State University in the USA, and in Germany is the GMD group. Much attention is paid to the application of KDD methods in biology and medicine. The greatest development was received by the applications of KDD associated with molecular biology, namely, with the decoding of macromolecules, and with the creation of new drugs. We should mention such companies as Base4 Bioinformatics, Bio Discovery, DNA Star, Molecular Simulations, Anvil Informatics, Bioreason, Cellomics, Incyte Pharmaceuticals. These data indicate that research in the field of KDD is currently experiencing rapid growth.

1. Basic methods of extracting knowledge from data

The development of computer technology has led to a significant increase in the volume of stored data, which, in turn, has complicated their analysis. At the same time, the need for such an analysis is quite obvious, because these data contain knowledge that can be used when making decisions.

Data Mining is a process of discovering in "raw data" previously unknown non-trivial practically useful and accessible interpretation of knowledge necessary for decision-making in various spheres of human activity. Data Mining is the revealing hidden patterns or relationships between variables in large amounts of raw data. It is subdivided into the tasks of classification, modeling and forecasting and others. The term "Data Mining" was introduced by Grigory Pyatetsky-Shapiro in 1989. The English term "Data Mining" does not have an

ISSN 2522-9818 (print)
Сучасний стан наукових досліджень та технологій в промисловості. 2020. № 3 (13) ISSN 2524-2296 (online)

unambiguous translation into Russian (data mining, data mining, information penetration, data / information extraction), therefore, in most cases it is used in the original. The most successful indirect translation is the term "intellectual analysis of data" (IAD).

IAD includes methods and models of statistical analysis and machine learning, distancing itself from them towards automatic data analysis. The IAD tools allow data analysis by subject specialists (analysts) who do not have the appropriate mathematical knowledge.

The information found in the process of applying Data Mining methods must be non-trivial and previously unknown, for example, average sales are not. Knowledge should describe new relationships between properties, predict the values of some features based on others, etc. The knowledge found should be applicable to new data with some degree of reliability. The usefulness lies in the fact that this knowledge can bring some benefit in its application.

Knowledge should be presented in a form that is understandable for a non-mathematical user. For example, the logical constructions "if ... then ..." are most easily perceived by humans. Moreover, such rules can be used in various DBMS as SQL queries. In the case where the extracted knowledge is not transparent to the user, there should be post-processing methods to bring it to an interpretable form. The algorithms used in Data Mining are computationally intensive. Previously, this was a limiting factor in the widespread practical application of Data Mining, but the increase in the performance of modern processors has removed the urgency of this problem. Now, within a reasonable time, it is possible to conduct a qualitative analysis of hundreds of thousands and millions of records [2].

To solve these problems, various Data Mining methods and algorithms are used. The following methods have gained great popularity: neural networks, decision trees, clustering algorithms, including scalable ones, algorithms for detecting associative links between events, etc.

Data analysis is based on modeling. Model building is a versatile way to explore the world around you. Modeling allows you to discover dependencies, extract new knowledge, predict, manage, and solve many other problems. Most economic systems are classified as complex, i.e. with a lot of elements and complex connections. The knowledge extraction technique is shown in fig. 1.

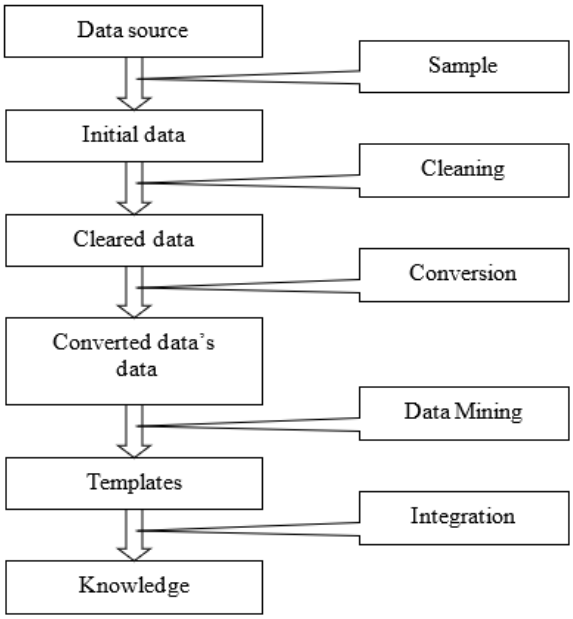


Fig. 1. Knowledge extraction technique

Despite the large number of various data analysis tasks, almost all of them are solved using a single methodology. It does not describe a specific algorithm or mathematical apparatus, but a sequence of actions that must be performed to build a model (extract knowledge). This technique does not depend on the subject area, it is a set of atomic operations, combining which you can get the desired solution [3, 4].

2. Tasks solved by Data Mining methods:

1. Classification is the assignment of objects (observations, events) to one of the previously known classes. The simplest and most common Data Mining task. As a result of solving the classification problem, features are found that characterize groups of objects of the studied data set – classes; on these grounds, a new object can be attributed to one or another class. The following methods can be used to solve the classification problem: Nearest Neighbor; k-Nearest Neighbor; Bayesian Networks; induction of decision trees; neural networks.

2. Regression, including forecasting problems. Establishing the dependence of continuous output on input variables. In regression and classification problems, it is required to determine the value of the dependent variable of an object based on the values of other variables that characterize this object.

Let there be given a finite set of objects $I = \{i_1, i_2, ..., i_j, i_n\}$. Each of the objects is characterized

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative Technologies and Scientific Solutions for Industries. 2020. No. 3 (13)

by some characteristic description $(x_1, x_2, ..., x_k, ..., x_m, x_{m+1})$. Let the feature values $(x_1, x_2, ..., x_k, ..., x_m)$ are known. Then the task is to determine the unknown feature x_{m+1}. If its set of values is finite, then the problem is called classification, and if countable or has the cardinality of the continuum, then we speak of the regression problem.

3. Clustering is a grouping of objects (observations, events) based on data (properties) that describe the essence of these objects. Objects within a cluster must be "similar" to each other and differ from objects included in other clusters. The more similar objects within a cluster and the more differences between clusters, the more accurate the clustering. Clustering is a logical continuation of the idea of classification. This task is more complicated; the peculiarity of clustering is that the classes of objects are not initially predefined. The result of clustering is the division of objects into groups.

4. Association – identifying patterns between related events. An example of such a pattern is the rule that indicates that event Y follows from event X. Such rules are called associative. This problem was first proposed to find typical shopping patterns in supermarkets, therefore it is sometimes also called market basket analysis. In the course of solving the problem of finding association rules, patterns are found between related events in the dataset. The difference between the association and the previous Data Mining tasks is that the search for patterns is carried out not based on the properties of the analyzed object, but between several events that occur simultaneously. The most famous algorithm for solving the problem of finding associative rules is the Apriori algorithm [1, 5].

5. Sequential Patterns – Establishing patterns between time-related events. Consistency allows you to find temporal patterns between transactions. The task of a sequence is like an association, but its purpose is to establish patterns not between simultaneously occurring events, but between events related in time (i.e., occurring at a certain interval in time). In other words, the sequence is determined by the high probability of a chain of events related in time. In fact, an association is a special case of a sequence with a time lag of zero. This Data Mining problem is also called the sequential pattern problem. The rule of sequence: after event X, event Y will occur after a certain time. The solution to this problem is widely used in marketing and management, for example, in managing the customer cycle (Customer Lifecycle Management).

6. Deviation analysis – identifying the most unusual patterns. Business analysis problems are formulated in a different way, but the solution of most of them comes down to one or another Data Mining problem or a combination of them. For example, risk assessment is a solution to a regression or classification problem, market segmentation is clustering, demand stimulation is associative rules. In fact, Data Mining tasks are elements from which you can assemble a solution to the vast majority of real business problems.

The potential of Data Mining provides opportunities for expanding the boundaries of technology application:

- allocation of types of subject areas with their corresponding heuristics, the formalization of which will facilitate the solution of the corresponding Data Mining problems related to these areas
- creation of formal languages and logical means, with the help of which reasoning will be formalized, and the automation of which will become a tool for solving Data Mining problems in specific subject areas;
- creation of Data Mining methods that can not only extract patterns from data, but also form some theories based on empirical data;
- overcoming the significant lag of the capabilities of Data Mining tools from theoretical achievements in this area.

If we consider the future of Data Mining in the short term, then it is obvious that the development of this technology is most related to business. In the short term, products of this class can become as common and necessary as e-mail and, for example, be used by users to find the lowest prices for a particular product or the cheapest tickets. In the long term, the future of Data Mining is truly exciting and it can be a search by intelligent agents, both for new treatments for various diseases, and a new understanding of the nature of the universe. However, Data Mining is fraught with potential danger and after all, an increasing amount of information becomes available through the worldwide network, including private information, and more and more knowledge can be extracted from it.

Studies note that there are both successful solutions using Data Mining and bad experiences with this technology. The areas where the application of this technology will be successful have the following features:

- require knowledge-based solutions;
- have a changing environment;
- have accessible, sufficient and meaningful data;
- provide high dividends from the right decisions.

To solve the above tasks, various Data Mining methods and algorithms are used. In view of the fact that Data Mining has developed and develops at the intersection of disciplines such as statistics, information theory, machine learning, database theory, it is quite natural that most Data Mining algorithms and methods were developed on the basis of various methods from these disciplines. For example, the k-means clustering procedure was simply borrowed from statistics. The following Data Mining methods have gained great popularity: neural networks, decision trees, clustering algorithms, including scalable ones, algorithms for detecting associative links between events, etc.

The use of Data Mining methods is actually the only way to benefit from the accumulated information, otherwise the collected data will be "dead weight". Data Mining allows you to extract knowledge from data and turn it into competitive advantages: predict qualitatively, more accurately identify target audiences, predict the development of events, manage risks, and more. Let's consider in detail the main components of the technology and methods of Data Mining.

ISSN 2522-9818 (print)
Сучасний стан наукових досліджень та технологій в промисловості. 2020. № 3 (13) ISSN 2524-2296 (online)

2.1 Data warehouses

The concept of a data warehouse is based on the idea of separating data used for operational processing and for solving analysis problems. This separation allows you to optimize both the online storage data structures for performing input, modification, deletion, and search operations, and the data structures used for analysis (for performing analytical queries). Different operational data sources (management systems) may contain data describing the same subject area from different points of view (accounting, warehouse accounting, planning department, etc.). A decision made on the basis of only one point of view may be ineffective or incorrect. Data warehouses allow you to integrate information that reflects different points of view into one subject area. Operational data sources tend to be developed at different times using different toolkits. This leads to the fact that the same objects are described in different ways. Integration of data into the warehouse solves this problem by bringing data to a single format [6–8].

Requirements for online data sources impose a limitation on the storage time in them, that is, those data that are not needed for online processing can be deleted from the database to reduce the amount of occupied resources. Analysis requires data for as long as possible. Unlike databases, in storages, data is only read after loading, which can significantly increase the speed of data access. Complex analytical queries against online data sources consume a large amount of resources on the computers on which they run. This leads to a decrease in system performance, which is unacceptable, since the execution time of operations in such systems is often very critical [9].

Thus, data prepared in a certain way and collected in storages can be used for analysis and decision-making based on them. Analysis subsystems (OLAP, Data Mining) are responsible for the formation of analytical queries to data and presentation of the results of their execution in decision support systems (DSS) [10].

2.2 Online data analysis systems

To analyze information, the most convenient way to represent it is a multidimensional model or a hypercube whose edges are dimensions. This allows you to analyze data from several dimensions at once, i.e. perform multivariate analysis. The concept of multivariate data analysis is closely related to operational analysis, which is performed using OLAP systems.

OLAP is a technology of online analytical data processing that uses methods and tools for collecting, storing and analyzing multidimensional data in order to support decision-making processes. The main purpose of OLAP systems is to support analytical activities, arbitrary requests from analyst users. The purpose of OLAP analysis is to test emerging hypotheses.

OLAP systems provide analysts with a means of testing hypotheses when analyzing data, that is, the main task of an analyst is to generate hypotheses based on their knowledge and experience.

By the type of database used, OLAP systems can be divided into several classes depending on the data storage structure:

- systems for operational analytical processing of multidimensional databases (or MOLAP systems), in which data are organized in the form of ordered multidimensional arrays of hypercubes or polycubes;
- systems of on-line analytical processing of relational databases (or ROLAP-systems), which allow presenting data in a multidimensional form, ensuring the transformation of information into a multidimensional model through an intermediate layer of metadata;
- hybrid systems of on-line analytical data processing (HOLAP-systems) are designed to combine the advantages and minimize the disadvantages of previous systems. They combine the analytical flexibility and speed of response of MOLAP systems with constant access to real data inherent in ROLAP systems.

The main disadvantages of the methods of operational data analysis discussed above are that in practice, the multidimensionality of data, as a rule, ends with the use of functions for extracting extensions from attributes of the "date / time" type: year, half year, quarter, month, week. All further actions are reduced to the sequence of execution of specialized queries: "selection" and "grouping" using the functions "CONT", "SUM", "AVG" or other similar.

2.3 Statistical methods for knowledge extracting

Domain-specific analytical systems are very diverse. The broadest subclass of such systems, which has become widespread in the field of financial market research, is called "technical analysis". It is a combination of several dozen methods for forecasting price dynamics and choosing the optimal structure of an investment portfolio, based on various empirical models of market dynamics. These methods often use a simple statistical apparatus, but maximally take into account the specifics that have developed in their field (professional language, systems of various indices, etc.)

The latest versions of almost all known statistical packages include Data Mining elements along with traditional statistical methods. But the main attention in them is paid to classical methods correlation, regression, factor analysis, etc. The disadvantage of systems of this class is the requirement for special training of the user.

There is an even more serious fundamental drawback of statistical packages that limits their use in Data Mining. Most of the methods included in the packages are based on a statistical paradigm in which the main figures are the average characteristics of the sample. And these characteristics, when studying real complex life phenomena, are often fictitious values.

2.4 Neural network mining algorithms

Neural networks are a large class of systems, the architecture of which is analogous to the construction of neural tissue from neurons. In one of the most common architectures, a multilayer perceptron with back propagation of an error, the operation of neurons as part of a hierarchical network is imitated, where each neuron of a

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative Technologies and Scientific Solutions for Industries. 2020. No. 3 (13)

higher level is connected by its inputs to the outputs of neurons of the underlying layer. The values of the input parameters are fed to the neurons of the lowest layer, on the basis of which it is necessary to make some decisions, predict the development of the situation, etc. These values are considered as signals transmitted to the next layer, weakening or amplifying depending on the numerical values (weights) attributed to interneural connections. As a result, at the output of the neuron of the uppermost layer, a certain value is generated, which is considered as a response and the reaction of the entire network to the entered values of the input parameters. In order for the network to be used in the future, it must be "trained" on the data obtained earlier, for which both the values of the input parameters and the correct answers to them are known. The training consists in the selection of the weights of interneuronal connections, ensuring the closest proximity of the network's responses to the known correct answers.

The main disadvantage of the neural network paradigm is the need to have a very large training sample. Another significant drawback is that even a trained neural network is a "black box". Knowledge, recorded as the weights of several hundred inter-neural connections, is completely beyond human analysis and interpretation.

2.5 Decision trees

Decision trees are a way of representing rules in a hierarchical, sequential structure, where each object has a single node that provides a solution. A rule is understood as a logical structure represented in the form "if ... then ...".

The field of application of decision trees is currently wide, but all the problems solved by this apparatus can be combined into the following two classes:

- Description of data. Decision trees allow us to store information about data in a compact form, instead of them we can store a decision tree that contains an accurate description of objects.

- Classification. Decision trees do an excellent job of classification tasks, i.e. assignment of objects to one of the previously known classes. The target variable must have discrete values. If the target variable has continuous values, decision trees allow you to establish the dependence of the target variable on independent (input) variables. For example, this class includes problems of numerical forecasting (prediction of the values of the target variable).

Let some training set *T* be given, containing objects (examples), each of which is characterized by m attributes, and one of them indicates that the object belongs to a certain class. The idea of constructing decision trees from a set *T*, first proposed by Hunt, is given by R. Quinlan.

Let $\{C_1, C_2, ..., C_k\}$ denote classes (class label values), then there are 3 situations:

- the set *T* contains one or more examples belonging to the same class C_k. Then the decision tree for *T* is a leaf that defines the class C_k;

- the set *T* contains no examples, i.e. empty set. Then it is again a leaf, and the class associated with the leaf is selected from another set other than *T*, for example, from the set associated with the parent;

- the set *T* contains examples belonging to different classes. In this case, the set *T* should be split into some subsets. For this, one of the features is selected that has two or more different values $O_1, O_2, ..., O_n$. T is split into subsets $T_1, T_2, ..., T_n$, where each subset Ti contains all examples that have the Oi value for the selected feature. This procedure will recursively continue until the finite set consists of examples belonging to the same class.

The above procedure is the basis of many modern algorithms for constructing decision trees, this method is also known as divide and conquer. Obviously, when using this technique, the decision tree will be built from top to bottom. Since all objects have been previously assigned to the classes we know, this process of building a decision tree is called supervised learning. The learning process is also called inductive learning or tree induction.

Today there are a significant number of algorithms that implement decision trees CART, C4.5, NewId, IT rule, CHAID, CN2, etc. But the most widespread and popular are the following two:

- CART (Classification and Regression Tree) is an algorithm for constructing a binary decision tree – a dichotomous classification model. Each node of the tree, when split, has only two children. As the name of the algorithm suggests, it solves classification and regression problems.

- C4.5 - an algorithm for constructing a decision tree, the number of children of a node is not limited. Can't work with a continuous target field, so only solves classification problems.

Most of the known algorithms are "greedy algorithms". If an attribute has been selected once and has been subdivided into subsets, the algorithm cannot go back and choose another attribute that would give a better subset. And therefore, at the stage of construction, it cannot be said whether the selected attribute will ultimately give the optimal partition.

Having considered the main problems that arise when constructing trees, consider their advantages: fast learning process; generation of rules in areas where it is difficult for an expert to formalize his knowledge; extraction of rules in natural language; intuitive classification model; high forecast accuracy, comparable to other methods (statistics, neural networks); construction of nonparametric models. Decision trees are an effective tool in decision support systems, data mining. Many data mining packages already include methods for constructing decision trees. In areas where the cost of error is high, they serve as an excellent support for the analyst or manager. Decision trees are successfully used to solve practical problems in the following areas:

- banking: assessing the creditworthiness of the bank's clients when issuing loans;

- industry: product quality control (detection of defects), non-destructive testing (for example, welding quality control), etc.;

- medicine: diagnostics of various diseases;

ISSN 2522-9818 (print)
Сучасний стан наукових досліджень та технологій в промисловості. 2020. № 3 (13) ISSN 2524-2296 (online)

- molecular biology: analysis of the structure of amino acids.

This is not a complete list of areas of use for decision trees. Many potential applications are not yet explored.

2.6 Limited search algorithms

Restricted search algorithms were proposed in the mid-60s by M.M. Bongard to search for logical patterns in data. Since then, they have demonstrated their effectiveness in solving a variety of problems from a wide variety of fields. These algorithms calculate the frequencies of combinations of simple logical events in subgroups of data. Examples of simple logic events: $X = a$; $X < a$; $X > a$; $a < X < b$ and others, where X is any parameter, a and b are constants.

One of the dangers in constructing a description is the possibility of the emergence of so-called prejudices, i.e. selection of conditions that satisfy the requirements set only on the training set and are useless outside of it. Apparently, the first to draw attention to this problem was M.M. Bongard. It is clear that by increasing the thresholds for the probability of a type I error and decreasing the probabilities of a type II error, we reduce the likelihood of prejudice. However, the problem of choosing thresholds that guarantee that the probability of the appearance of even one prejudice does not exceed a given value remains relevant.

2.7 Reasoning methods based on similar cases

The idea of Case Based Reasoning (CBR) is simple at first glance. In order to make a forecast for the future or choose the right decision, these systems find in the past close analogs of the current situation and choose the same answer that was correct for them. Therefore, this method is also called the "nearestneighbor" method.

Recently, the term memory based reasoning has also become widespread, which focuses on the fact that a decision is made on the basis of all the information accumulated in memory. CBR systems perform well in a wide variety of tasks. Their main disadvantage is that they do not create any models or rules generalizing previous experience at all. In choosing a solution, they are based on the entire array of available historical data, so it is impossible to say on the basis of which specific factors CBR systems base their answers. Another disadvantage is the arbitrariness that CBR systems allow when choosing a measure of "proximity". This measure most decisively determines the size of the set of use cases that must be stored in memory in order to achieve a satisfactory classification or forecast.

2.8 Genetic algorithms

Genetic algorithms are designed to solve optimization problems. An example of such a task is training a neural network, that is, the selection of such values of the weights at which the minimum error is achieved. Moreover, the genetic algorithm is based on a random search method. The main disadvantage of random search is that we don't know how long it will take to solve the problem. To avoid such a waste of time in solving the problem, methods are used that have manifested themselves in biology, methods discovered in the study of the evolution and origin of species. As you know, in the process of evolution, the fittest individuals survive. This leads to the fact that the fitness of the population increases, allowing it to better survive in changing conditions. This algorithm was first proposed in 1975 by John Holland at the University of Michigan. It was named "Holland's reproductive plan" and formed the basis for almost all variants of genetic algorithms.

It is known from biology that any organism can be represented by its phenotype, which actually determines what an object is in the real world, and a genotype, which contains all information about an object at the level of the chromosome set.

Moreover, each gene, that is, an element of genotype information, is reflected in the phenotype. Thus, to solve problems, we need to present each feature of an object in a form suitable for use in a genetic algorithm. All further functioning of the mechanisms of the genetic algorithm is carried out at the level of the genotype, making it possible to do without information about the internal structure of an object, which determines its widespread use in a variety of tasks. The most common form of genetic algorithm uses bit strings to represent the genotype of an object. Moreover, each attribute of an object in the phenotype corresponds to one gene in the object's genotype. A gene is a bit string, most often of a fixed length, that represents the value of this trait.

As you know, in the theory of evolution, an important role is played by how the traits of parents are transmitted to descendants. In genetic algorithms, the crossing operator (also called crossover or crossing over) is responsible for the transmission of parental traits to offspring. This operator defines the transfer of traits from parents to descendants. He acts as follows: two individuals are selected from the population who will be the parents; the break point is determined (usually randomly); a child is defined as the concatenation of part of the first and second parent. For the functioning of the genetic algorithm, these two genetic operators are sufficient, but in practice, some additional operators or modifications of these two operators are also used. For example, a crossover may not be single-point (as described above), but multi-point, when several break points are formed (most often two). In addition, in some implementations of the algorithm, the mutation operator is the inverse of only one randomly selected bit of the chromosome.

2.9 Systems for visualizing multidimensional data

Tools for visualization of multidimensional data are supported by all Data Mining systems. At the same time, a very impressive market share is occupied by systems that specialize exclusively in this function. In such systems, the main attention is focused on the friendliness of the user interface, which makes it possible to associate various parameters of the scatter diagram of database objects (records) with the analyzed indicators. These parameters include color, shape, orientation relative to its own axis, dimensions and other properties of graphic elements of the image. In addition, data visualization

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative Technologies and Scientific Solutions for Industries. 2020. No. 3 (13)

systems are equipped with convenient tools for scaling and rotating images.

2.10 Fuzzy logic methods

The mathematical theory of fuzzy sets and fuzzy logic are generalizations of classical set theory and classical formal logic. These concepts were first proposed by the American scientist Lotfi Zadeh in 1965. The main reason for the emergence of a new theory was the presence of fuzzy and approximate reasoning when a person describes processes, systems, objects. Before the fuzzy approach to modeling complex systems was recognized all over the world, more than one decade passed since the inception of the theory of fuzzy sets. And on this path of development of fuzzy systems, it is customary to distinguish three periods.

The first period (late 60s – early 70s) is characterized by the development of the theoretical apparatus of fuzzy sets (L. Zade, E. Mamdani, Bellman).

In the second period (70s - 80s), the first practical results appear in the field of fuzzy control of complex technical systems (a steam generator with fuzzy control). At the same time, attention began to be paid to the issues of constructing expert systems based on fuzzy logic, the development of fuzzy controllers. Fuzzy expert systems for decision support are widely used in medicine and economics.

In the third period, which lasts from the end of the 80s and continues at the present time, software packages for building fuzzy expert systems appear, and the fields of application of fuzzy logic are significantly expanding. It is used in the automotive, aerospace, transportation, home appliance, finance, analysis and management industries, and many others.

The triumphal march of fuzzy logic around the world began after Bartholomew Kosco proved the famous FAT (Fuzzy Approximation Theorem) in the late 80s. In business and finance, fuzzy logic gained acceptance after in 1988 a fuzzy rule-based expert system for predicting financial indicators was the only one predicting a stock market crash. And the number of successful fuzzy applications is currently in the thousands [11–14].

2.11 The task of finding association rules in datasets

Association rules allow you to find patterns between related events. The first association rule search algorithm, called AIS, was developed in 1993 by researchers at the IBM Almaden Research Center. The mid-90s of the last century saw the peak of research work in this area, and since then, several algorithms have appeared every year. [15, 16].

Subject areas in which the method of assessing the associative properties of data is most often used:

- retail: identifying products that are worth promoting together; selection of the location of the product in the store; analysis of the consumer basket; demand forecasting;
- marketing: search for market segments, trends in consumer behavior;
- customer segmentation: identifying the general characteristics of the company's customers, identifying groups of buyers;
- catalog design, analysis of sales campaigns, determination of customer purchase sequences (which purchase will follow the purchase of product A);
- analysis of web logs.

Finding association rules is one of the main approaches to data mining. The search reveals hidden connections of seemingly unrelated data. These connections are the rules. Those that exceed a certain threshold are considered interesting. One of the most frequently cited examples of the search for association rules is the problem of finding stable links in the shopping cart (Market-Basket Problem) [1, 17].

The challenge in finding lasting links in a shopping cart is to determine which items are being purchased by customers together, so that marketers can properly place those items in the store to increase sales, as well as make other decisions to drive sales. It is the ability to detect hidden rules that makes the search for association rules valuable and conducive to the search for knowledge [18–20].

3. Let's formulate the problem of data mining in relational systems.

By the task of data mining we mean the procedure for finding all pairs of attributes of a relational database that satisfy the condition for which group operations are executable.

The problem can be solved in at least two ways:

1. Based on the rules of normalization theory, analyzing the second and third normal forms, select all pairs of functionally dependent attributes $X \to Y$ from the relation $R\left(K_1, K_2, A_1, ..., A_n, B_1, ..., B_m\right)$. The list of pairs will also include attributes that are transitively dependent among themselves through relations in which a composite key is used as key attributes.

2. Based on an analysis of the database schema normalized to third normal form.

3.1 Relational database structural specification

To construct a structural diagram of databases, traditional means of specifying a relational data model are used [21-25]. The main structural unit of data in the relational model is an n-ary relation, which is a finite subset of the Cartesian product of domains, that is, sets of atomic values of data elements – the attributes of the relation.

Let R be a finite set of database relationship names;

D – $\{D_1, ..., D_i\}$ – a set of domains, where every D_i domain is a named set of atomic values of data elements;

A – a finite set of names of attributes of relations;

dom – mapping from A to D, defining from which domain the attribute values are selected.

A pair $\langle A_i, domA_i \rangle$ where $A_i \in A$ is called an attribute. The structural scheme S_i of a relationship

ISSN 2522-9818 (print)
Сучасний стан наукових досліджень та технологій в промисловості. 2020. № 3 (13) ISSN 2524-2296 (online)

R_i $(R_i \in R)$ can be represented in a form $R_i(A_1,...,A_n)$ where all A_i are different. A relationship r_i can be defined as an extension of a schema $S_i : r_i \subseteq domA_1 \times ... \times domA_n$. Permutation of attributes in a schema does not generate a new extension, and a set $\{A_1,...A_n\}$ of relationship attributes R_i sets the type of relationship. An expression $R_i = A_1...A_n$ is used to specify the media composition. The U block diagram of a relational database is a specification of the form $(R_1,...,R_p)$, where $R_i \in R$ and all R_i are different.

Conceptually, a relational base is an information-logical model of a certain subject area, such that each extension corresponds to a certain state of this area at a certain moment in the discrete current time. ach state is modeled by an ordered set of data item values corresponding to the values of properties of objects in the domain.

An object of a specific type corresponds to a tuple of a relation of a specific type. Note that the relational data model assumes a strong typing of objects, the use of well-defined categories, such as the type of an object, an attribute (property) of an object, a domain, and the assignment of each value and an ordered set of values to one of these categories. Objects of a certain type have a certain set of properties, which is specified in the relational model by a relationship schema [26-29].

At the design stage, the developer performs a mandatory procedure - normalizing the relational database schema to one of the normal forms, as a rule, this is the third normal form (3NF).

Aggregate functions are functions that determine the number of records in a table, count the number of values in a column or find the minimum and maximum values for it, and also sum data. Aggregate functions include functions COUNT, SUM, MAX, MIN, AVG and possibly others suggested by the developer.

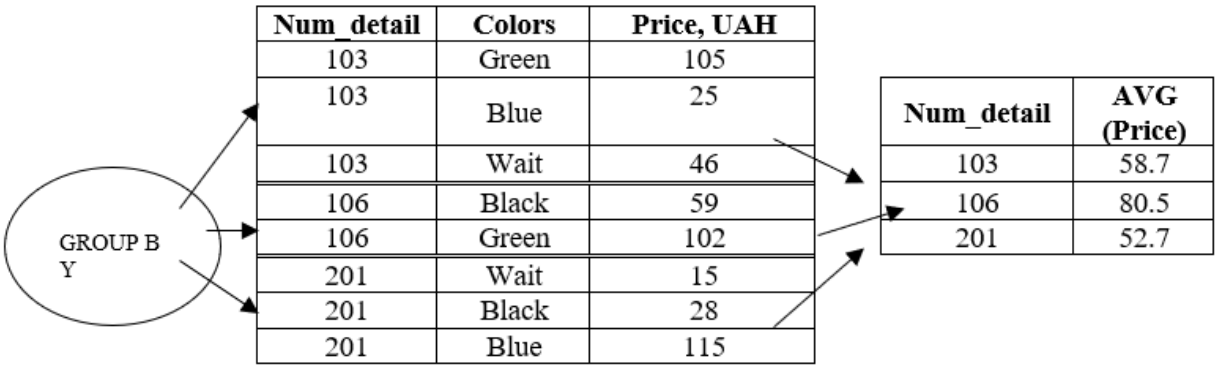
To apply aggregate functions in calculations with respect to some group of identical values, the Group By parameter is used. This parameter "compresses" the same values for the specified attribute into a single row of totals. To find the average price for a part, you can formulate a query in SQL.

SELECT Num_detail, AVG(Price) FROM Table GROUP BY Num_detail

Example.

The problem of finding associative dependencies in a relational database and its solution.

Let's consider the technology for finding associative data dependencies using the following example. For the analysis, the "DETAILS" relational relation with the following scheme <Num_detail, Colors, Price> is presented. The example shows that there is a correspondence between two information units: part number and price. Thus, it is possible to formulate a production expression of the form {Price ≠∅; Num_detail⇒ AVG(Price)}, where Num_detail ⇒ AVG(Price) –is a product core, Price ⇒∅ is a product core applicability condition; if the logical expression of the condition evaluates to "true", then the production core is activated. The diagram of using the AVG aggregate function relative to the grouped data is shown in fig. 2.



Num_detail	Colors	Price, UAH
103	Green	105
103	Blue	25
103	Wait	46
106	Black	59
106	Green	102
201	Wait	15
201	Black	28
201	Blue	115

Num_detail	AVG (Price)
103	58.7
106	80.5
201	52.7

Fig. 2. Query with grouping

When constructing production rules, the query syntax is used, which defines the structure of the rule core. The syntactic structure, in turn, is fixed and has the form: the left side of the rule corresponds to the attribute (attributes) by which the grouping is carried out, and the right side corresponds to the attribute (s) to which the aggregate function is applied. If a set of queries is used, consolidated, for example, in a transaction, then they form a production system. In the product system, special product management procedures must be specified, with the help of which the kernel is updated [30, 31].

Statement:

Let there be a relational relation for which a set of transactions is implemented $T = \{T_1,..., T_n\}$, $D = \{d_1,..., d_n\}$ – the set of elements of which the transaction consists of *T*, that is $T_i \subseteq D$ и Ω = {COUNT, SUM, MAX, MIN, AVG,…} – a set of aggregate functions. Each transaction is a binary vector, where $T_i = 1$, if the d_i element is present in the transaction and $T_i = 0$ otherwise. Transaction T_i has a set of elements $X \subseteq D$, if $X \subset T_i$. Then products of the form $\{P; X \Rightarrow \varpi(Y)\}$,where *P* – the condition for activating the

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative Technologies and Scientific Solutions for Industries. 2020. No. 3 (13)

core of a rule will be called a functional associative rule if $X \subset D$, $Y \subset D$, $X \cap Y = \varnothing$ и $\varpi \in \Omega$.

The purpose of data analysis is to establish the following dependencies: if a certain set of *X* elements was encountered in a transaction, then, based on this, we can conclude that another set of *Y* elements should also appear in this transaction. Establishing such dependencies makes it possible to find simple and intuitive rules [32, 33].

In general, the formation of association rules can be represented in two stages:

- selection of all the necessary sets of elements;
- generation of rules from sets of elements using the required functions.

As a means of displaying functional association rules, we use the semantic network (SN): $S = (T, \Omega)$, where *T* – a set of transactions, the elements of which act as nodes of the SN, Ω – set of aggregate functions representing the relationship between vertices (arcs of a graph).

Let a system of functional association rules be given:

{Quantity ≠∅; Product ⇒ SUM(Quantity)
Price ≠∅; Product ⇒ MIN(Price)
Product ≠∅; Name_Customer⇒ COUNT(Product)
Quantity ≠∅; Name_Customer, Product ⇒ MAX(Quantity)
Price ≠∅; Material ⇒ MIN(Price)
Quantity ≠∅; Price ⇒ MIN(Quantity)}

For the system presented above, we construct a semantic network *S*, shown in fig. 3. To identify the constituent left parts of the core of the rule in the network, the notation $\varpi^i \left(i = \overline{1, \infty}\right)$ is used, where *i* – shows the ratio of the constituent vertices.

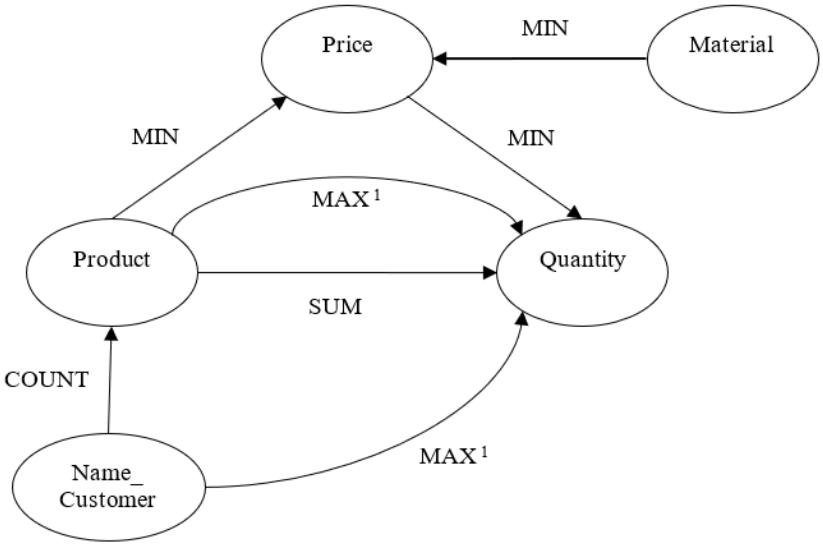


Fig. 3. Semantic network of the system of functional association rules

Using the structure of the SN, consider the inverse problem. Let's formulate a query based on the analysis of paths in the SN. For example, what information can be obtained from data on the names of suppliers (Name_Customer). Such a request can be formulated as follows:

SELECT Name_Customer, COUNT(Product) FROM Table
GROUP BY Name_Customer

SN analysis can expand the knowledge base. Having determined the necessary path in the network, the request can be composed, for example, as follows:

SELECT Name_Customer, Product, MIN(Price) FROM Table
GROUP BY Name_Customer, Product

Thus, from the expressions "if there is a supplier, then you can find the number of his deliveries" and "if there are supplies", then you can find their minimum cost, "if there is a supplier and his supplies", then you can find the minimum cost of these supplies [34].

Conclusions

In integrated information systems that combine a relational database and a decision support subsystem, various tasks are solved, including those related to the formalization and representation of knowledge. An important role in the search for knowledge is played by the structural characteristics of the source from which the intelligent system can obtain knowledge. Thus, it is necessary to have tools with which it is possible to acquire knowledge for an intelligent subsystem.

The main source of knowledge is the relational database. In this regard, the study of data properties is an urgent task in the construction of systems of association rules. On the one hand, associative rules are close to logical models, which makes it possible to organize efficient inference procedures on them, and on the other hand, they more clearly reflect knowledge than classical models. They do not have the strict limitations typical of

ISSN 2522-9818 (print)
Сучасний стан наукових досліджень та технологій в промисловості. 2020. № 3 (13) ISSN 2524-2296 (online)

logical calculus, which makes it possible to change the interpretation of product elements.

The search for association rules is far from a trivial task, as it might seem at first glance. One of the problems is the algorithmic complexity of finding frequent itemsets, since as the number of items grows, the number of potential itemsets grows exponentially.

The article deals with the analysis of a relational database in order to identify production rules, based on some structural specifications of the relational data model. One class of products is considered, namely functional association rules, which allows, using the properties of keys, primary and external, to generate a system of functional association rules and build on their basis a semantic network that allows you to expand (deduce) a given set of rules. Further research of the search and analysis procedures for products can be directed towards the development of methods for inference of association rules based on a certain system of axioms. In addition, when searching for rules, you can use other properties of the relational data model, such as functional and other types of dependencies, as well as some properties of the language specification, in particular the mathematical apparatus of relational calculus.

References

1. Gavrilova, T. A., Khoroshevsky, V. F. (2000), *Knowledge Base of Intellectual Systems*, SPb: Peter, 384 p.
2. Yesin, V. I. (2012), "Reinzhynirynh isnuyuchykh baz danykh", *Systemy obrobky ynformatsyy*, KHNU im. V.N. Karazina, Kharkiv, Vol. 2, No. 3 (101), P. 188–191.
3. Borisov, A. N., Alekseev, A. V., Merkur'eva, G. V. (1989), "Processing of fuzzy information in decision-making systems", *Radio and communication.*
4. Ed. S. Osugi, Y. Saeki (1990), *The acquisition of knowledge*, Moscow, Mir, 304 p.
5. Date, K. (2001), *Introduction to database systems* : trans. from English, Moscow, Publishing House "Williams", 1072 p.
6. Filatov, V., Rudenko, D. Grinyova, E. (2014), "Means of integration of heterogeneous data corporate information and telecommunication systems", *Proceedings of the 24th International Crimean Conference Microwave and Telecommunication Technology (CriMiCo-2014), 7-13 sept. 2014*, Sevastopol, Ukraine, P. 399–400.
7. Fillmore, C. J. (1978), *The case for case*, Universals in linguistic theory, N. Y., Holt, Rinehart and Winston Inc., 234 p.
8. Glava, M., Malakhov, V. (2018), "Information Systems Reengineering Approach Based on the Model of Information Systems Domains", *International Journal of Software Engineering and Computer Systems (IJSECS)*, University Malaysia Pahang, Vol. 4, P. 95–105. DOI: 10.15282/ijsecs.4.1.2018.8.0041
9. Avrunin, O. G., Bodianskyi, Ye. V., Kalashnyk, M. V., Semenets, V. V., Filatov, V. O. (2018), *Suchasni intelektualni tekhnolohii funktsionalnoi medychnoi diahnostyky*, KhNURE, Kharkiv, 236 p. DOI: 10.30837/978-966-659-236-4
10. Kosenko, V. (2017), "Principles and structure of the methodology of risk-adaptive management of parameters of information and telecommunication networks of critical application systems", *Innovative Technologies and Scientific Solutions for Industries*, No. 1 (1), P. 46–52. DOI: https://doi.org/10.30837/2522-9818.2017.1.046
11. Zade, L. A. (1976), *The concept of a linguistic variable and its application to making approximate decisions*, Moscow, Mir, 165 p.
12. Asai, K., Vatada, D., Iwai, S. et al. (1993), *Applied fuzzy systems*, Ed. T. Terano, C. Asai, M. Sugeno, Moscow, Mir, 368 p.
13. Zadeh, L. A. (1974), "Basics of a new approach to the analysis of complex systems and decision-making processes", *Math Today*, Moscow, Znanie, P. 5–49.
14. Kent, W. (1981), "Consequences of assuming a universal relation", *ACM Trans. on Database Systems*, Vol. 3, P. 3–17.
15. Korneev, V. V., Gareev, A. F., Vasyutin, S. V., Reich, V. V.(2001), *Database*, Intellectual information processing, 2nd ed., Moscow, Nolidge, 496 p.
16. Dubois, D., Prades, A. (1990), *Theory of opportunities. Applications to the representation of knowledge in computer science*, Moscow, Radio and communication, 288 p.
17. Sichkarenko, V. A. (2002), *SQL 99 Database Developer Guide*, Moscow, DiaSoftUP, 816 p.
18. Rumbaugh, J., Blaha, M. (1991), *Object-Oriented Modeling and Design*, N. J., Prentice Hall, 348 p.
19. Schmid, H. A., Swenson, J. R. (1975), "On the semantics of the relation model", *Proc. of ACM SIGMOD Int. Conf. Management of Data*, P. 211–223.
20. Langefors, B. (1974), "Information systems", *Information Processing 74*, Amsterdam, North-Holland, P. 937–945.
21. McLeod, D. (1979), *The semantic data model*, MIT Press.
22. Tsalenko, M. Sh. (1989), *Modeling semantics in databases*, Moscow, Nauka, Main ed. ph.-mat.lit., 288p.
23. Schenk, R. (1980), *Processing Conceptual Information*, Moscow, Energy, 268 p.
24. Rob, P., Coronel, K. (2004), *Database Systems: Design, Implementation, and Management* : Trans. from English, SPb., BHV-Petersburg, 1023 p.
25. Langefors, B. (1980), "Infological models and information user views", *Inform. Systems*, Vol. 5, P. 17–32.
26. Buslik, M. M. (1993), *Optimal image of a real database* : Monograph, Kyiv, ISDO, 84 p.
27. Martin, J. (1980), *Database Organization in Computing Systems* : Tr.from English, Moscow, Mir, 662 p.
28. Maltsev, A. I. (1970), *Algebraic systems*, Moscow, Nauka, 392 p.
29. Cycritis, D., Lokhovsky, F. (1985), *Data Models* : Trans. from English, Moscow, Finance and Statistics, 344 p.
30. Filatov, V., Semenets, V. (2018), "Methods for Synthesis of Relational Data Model in Information Systems Reengineering Problems", *Proceedings of the International Scientific-Practical Conference "Problems of Infocommunications. Science and Technology" (PIC S&T-2018), 9-12 oct. 2018*, Kharkiv, Ukraine, P. 247–251.
31. Filatov, V., Kovalenko, A. (2020), "Fuzzy systems in data mining tasks", DOI: 10.1007/978-3-030-35480-0_6
32. Filatov, V., Radchenko, V. (2015), "Reengineering relational database on analysis functional dependent attribute", Proceedings of the X Intern. Scient. and Techn. Conf. "Computer Science & Information Technologies" (CSIT'2015), 14-17 sept. 2015, Lviv, Ukraine, P. 85–88.
33. Filatov, V. (2014), "Fuzzy models presentation and realization by means of relational systems", *Econtechmod: an international quarterly journal on economics in technology, new technologies and modelling processes*, Lublin, Rzeszow, Vol. 3, No. 3, P. 99–102.

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) *Innovative Technologies and Scientific Solutions for Industries. 2020. No. 3 (13)*

34. Filatov, V., Doskalenko, S. (2018), "The Approach to Searching for Functional Dependences of Data in Relational Systems", *Innovative Technologies and Scientific Solutions for Industries*, No. 3 (1), P. 54-58. DOI: https://doi.org/10.30837/2522-9818.2018.3.054.