Research on Data Mining Technology Based on Business Intelligence

Yang WANG
Communication University of China, China

Keywords: Data mining, Business intelligence, Clustering, Classification model.

Abstract. In this paper, the development status and trends of business intelligence was introduced, and the data mining techniques used in business intelligence such as classification and clustering methods were studied.

Introduction

Business intelligence is a term that firstly appeared in the foreign business communities in the late 1990s [1]. It is a series of methods, techniques and software used to improve the performance of the enterprise. It applied the advanced information technology to the entire enterprise, not only providing information acquisition ability, but also transforming it into the competitive advantage of the enterprise through the development of information, which was also known as intelligence in the chaos of the world. Therefore, more and more enterprises put forward their demand for BI, and make the BI as an effective means to help enterprises to achieve business objectives.

At present, business intelligence is usually understood as the transformation of the existing data in the enterprise to knowledge, to help enterprises to make informed business decision-making tools. Here to talk about the data from enterprise business system of orders, inventory, account transactions, customers and suppliers and enterprises from the industry and competitor data, and various data from other enterprises in which the external environment. And business intelligence can assist the business decision-making either the operation layer, or the management and strategic level of decision-making.

In order to transform data into knowledge, data warehouse, on-line analytical processing (OLAP) tools and data mining techniques are needed. Therefore, speaking of the technical level, business intelligence is not a new technology. It is just ETL, data warehouse, OLAP, data mining, data show and other comprehensive applications of technology.

The business intelligence should be seen as a solution appropriately. The key structure of business intelligence is to extract useful data and clear it from the many different business operation systems, with ensuring the correctness. After the process of extraction, transformation and load known as the ETL process, adding the consequences into an enterprise data warehouse to get a global view of the enterprise data. On the basis of warehouse results, with using the appropriate query and analysis tools, data mining tools, and OLAP tools, information data becomes knowledge aided decisions. At last, the knowledge is presented to the manager, which provides support for the manager's decision making process.

The development of the U.S. business intelligence industry may become a model of the domestic market. Currently, the U.S. business intelligence market is not only IBM, SAP, Oracle and other giants involved. And make some segments of the medium-sized companies access to moisture. For example, the business intelligence software development company GoodData recently announced the completion of the fourth round of financing, the scale of $22 million, the total amount of GoodData's total financing has reached $75 million 500 thousand. With IBM, SAP, Oracle and other giants to provide a different suite of solutions, GoodData will be all of its data analysis services have moved to the cloud, SaaS model to provide to the enterprise. Gooddata cloud platform, enterprises can company has import data, the data tracking, segmentation and visualization, analysis processing, Gooddata now has customers include software AG, Time Warner Cable and so on. And that's exactly what it is now. Related survey shows that a major demand of large overseas company's chief information officer (CIO) attention is business transformation,
which involves three key techniques are business intelligence analysis, mobile devices and cloud computing and applications.

At present, the main domestic commercial intelligence industry downstream markets involve the telecommunications, finance, energy, transportation and government whose level of informationize is relatively high in the field. The investment of these industries in the construction of IT systems maintained a relatively high growth rate especially in recent years, with large effective affects on business intelligence development.[2]

It can be expected that with the general increase of domestic enterprise informationize level, cost of commercial competition and human ascension will urge the industry started to increase investment in business intelligence, increasingly broaden the coverage of the downstream business intelligence. And business intelligence products and services of demands will maintain a steady growth.

On the current view, the telecommunications and financial industry is still the fastest growing business intelligence industry downstream customers. In some large telecom operators, business intelligence system has been widely used. And in some companies, business intelligence system has attracted more and more emphasis on functional areas of its IT architecture.

According to the planning for the future development of China's three major telecom operators of business intelligence (BI) software, the telecom operators will pay more attention to investment in software and it committed to the business intelligence applications into all aspects of business management and business. With the increasing competition of the three operators of the whole business, it needs more energy for marketing, customer management, customer retention and so on, the business intelligence system can effectively help operators to achieve the above objectives.

This paper is structured as follows. The second part is about the classification of data mining technology in business intelligence. The third part studies the classification and clustering of the main data mining techniques used in business intelligence. The last part proposes the conclusion.

### Classification of Data Mining Used in Business Intelligence

Due to the data mining tech derives from various subjects, the research of data mining is expected to generate a large number of data mining systems[3,4,5]. This classification can help users to distinguish the data mining systems, and to determine the most suitable data mining system for their needs. According to different standards, the data mining system can be classified as follows:

**Classified according to the Type of Database Mining**

Data mining system can be classified according to the type of database. The database system itself can be classified according to different standards such as data model, data type or application classification, each of which may need their own data mining technology. The data mining system can be classified correspondingly. For example, according to the classification of the data model, it can be related to the transaction, the object of a relational or data warehouse mining system. If a particular type of data is classified according to the processing data, the data mining system or the World Wide Web mining system can have the space, time series, text, streaming data, multimedia data mining system.

**Classified according to the Type of Knowledge Mining**

Data mining system can classified according to the mining of knowledge, namely according to the function of data mining, such as characterization, discrimination, association, correlation analysis, classification, prediction, clustering, outlier analysis and evolution analysis. An integrated data mining system typically provides multiple and or integrated data mining functions. In addition, the data mining system can be distinguished according to the granularity of the mining knowledge or the abstraction layer, including the high abstraction layer and the original data layer. An advanced data mining system should support the knowledge discovery of multi abstract layers. The data mining system can also be classified as the rule of mining data and the singularity of the data mining.
Classified according to the Type of Technology Used

Data mining systems can also be classified according to the data mining techniques used. These techniques can be described according to the level of user interaction such as automatic system, interactive exploration system, query driven system, or the data analysis methods for example database or for data warehouse technology, machine learning, statistics, visualization, pattern recognition, and neural network. Complex data mining systems usually adopt a variety of data mining techniques or the use of effective, integrated technology.

Classified according to the Applications

Data mining systems can also be classified according to their applications. For example, there may be some data mining system particularly suitable for financial, telecommunications, and the stock market. Different applications usually require particularly effective approaches. Therefore, the generalization of the universal data mining system may not be suitable for the specific domain of mining tasks.

The following focuses on the main data mining techniques used in several business intelligence.

Main Data Mining Techniques Used in Business Intelligence

Two main data mining technology used in business intelligence is introduced as follows:

Classification

The classification of data mining is already known training data feature extraction and classification results, to find a reasonable description or model for each class and use the classification description or model of the unknown new data classification. Classification is a very important method of data mining. The concept of classifier is to construct a classification function or construct a classification model based on the data. The function or model can map the data record in the database to a certain one in a given class so that it can be applied to the data prediction. To construct a classifier, a training sample data set is needed as the input. The training set consists of a set of database records or a tuple. Each record is a feature vector consisting of the value of the field. These fields are called attributes that are used to classify attributes called labels, which are labeled as training sets. A specific sample of the form can be expressed as \((V_1, V_2, ..., V_N; C)\) where \(V_i\) represents the field value, and the \(C\) represents the class. The training set is the basis of constructing classifier. The training set is a database table that contains a number of attributes, one of which is formulated as a category label. The type of label attributes must be discrete, and the number of possible values of the tag attribute is as little as possible, preferably two or three values. The less the number of tags is, the lower the error rate of the classifier is constructed. An algorithm for automatically creating a classifier from a training set is called a generator. After the classifier is generated, it can be used to classify the records of the data set, which is not included in the data set. Common classification techniques are decision tree classification, selection tree classification, Bias classification and neural network.

Clustering

As one of the main methods of data mining, clustering has attracted more and more attention. The clustering method is to gather a large number of data objects so that the same cluster within the object of the same similarity as the maximum, while the similarity of different clusters within the object to achieve the minimum. That is to say, after the formation of the clustering, objects in the same cluster have high similarity, meanwhile which does not belong to the clustering objects have vastly differences. Comparing the clustering with classification, classification algorithm analyses the known data sets, and the clustering algorithm is for the unknown data. The input of clustering is a set of unclassified records which did not know in advance to be divided into several categories. It is through the analysis of the data, according to the criterion of classification, rational division of a collection of records, in order to determine each record belongs to the category. Different clustering algorithm used to describe the similarity function is also different. Some use Euclidean distance or
Markov distance as well as some of the cosine of the angle of the vector, but some of the other methods of measurement. When the type number is not known in advance, or it is difficult to distinguish different types of class probability density function with parameter estimation and nonparametric estimation, the clustering analysis is needed. Some clustering analysis algorithms can be used to determine the number of types automatically, without the need to predict as the prerequisite, can also be given as the termination conditions of the algorithm. If not given, then how to determine automatically in the clustering process is a key problem in cluster analysis. Clustering is an important technique in data mining. It has been widely used in the fields of customer classification, gene recognition, text classification, spatial data processing and so on. Cluster analysis can help the market analysts to find different customer groups, and use the purchase model to characterize the characteristics of different customer groups. Cluster method includes statistical method, machine learning method, neural network method and database oriented method.

In statistical method, cluster is also called cluster analysis, which is one of the three methods of multivariate data analysis. The other two are regression analysis and discriminant analysis, mainly based on the geometric distance based clustering, such as Euclidean distance, etc. The traditional statistical clustering analysis method includes the system clustering method, the decomposition method, the adding method, the dynamic clustering method, the ordered sample clustering, the overlapping clustering and the fuzzy clustering and so on. Cluster analysis method does not have a linear computational complexity, and it is difficult to apply to a very large database.

In machine learning, clustering is called unsupervised or non- teacher induction. Compared with classification learning, clustering is not labeled, and it needs to be determined automatically by the clustering algorithm. A lot of artificial intelligence literature, clustering is also known as the concept of clustering, the distance is no longer statistical methods in the geometric distance, but according to the description of the concept to determine.

The clustering method can be divided into the following categories: partition method, hierarchical method, density based method, grid based method and model based method.

**Partition Method.** Given a database containing N data objects or tuple, the structure of K is divided, each partition represents a cluster (k-cnt). For a given K, the first process is to create K initial partition and through iterative steps to change division making each improvement after the partition scheme better than the previous. This clustering method is very good for the classification of small and medium size databases. In order to cluster large data sets and to deal with the complex shape of the cluster, the method based on the division needs to improve the expansion.

**Hierarchical Method.** Decomposition of a given set of data objects until a certain condition is satisfied. The hierarchical method is called as the two basic methods of aggregation and splitting. The method begins by combining each object as a single group, and then merging the similar objects or groups in succession until all the combinations are combined and a termination condition is reached. The split method, will begin with all the object placed in a cluster in iteration of each step, a cluster is split into smaller clusters until the final on each object in a single cluster or to a termination condition. The defect of hierarchical clustering method is that after each step is completed, the method cannot be revoked. Therefore it cannot correct the wrong decision. The improved hierarchical clustering method is a multi-stage clustering method.

**Density Based Method.** Density based clustering method is to find the clustering results of arbitrary shape. The main idea is to continue clustering as long as the density of the neighboring region exceeds a certain threshold, and the data points in the data sparse area are considered as noise data. This method can be used to filter the outlier data of “noise” and find the clustering of arbitrary shape.

**Grid Based Method.** The object space is quantized into a finite number of elements to form a grid structure. All the cluster operations are performed on this grid structure. The main advantage of this method is that its processing speed is very fast, and its processing time is independent of the number of data objects, which is only related to the number of cells in each dimension.

**Model Based Approach.** A model based approach to each cluster assumes a model, and then goes to find the data set that can be well satisfied with this model. Based on the model, it is possible
to locate the cluster by constructing the density function of the spatial distribution of the response data, and it is possible to automatically determine the number of clusters by using the standard statistics.

**Conclusion**

This paper introduces the development status and trends of business intelligence, the data mining techniques used in business intelligence classification and clustering methods were the main objects we studied.

**References**


