Research on the Application Scheme of Enterprise Data Mining Based on Extension Set

Jiang Chunxu, Guan Bo

Baicheng Normal College

Keywords: Extenics; extension set; matter-element analysis; data mining; enterprise information; data quality; data mining consulting

Abstract. The quality difference of data itself cause low credibility of data mining conclusions, which has become an important factor affecting the application of data mining. The cleansing algorithms and tolerate algorithms designed in view of incomplete data can not solve this problem fundamentally. By means of in-depth analysis of the reasons for this contradictory phenomenon, the matter-element extension set is established for corporate data, and the application scheme of data mining enterprise on basis of extension method has been proposed. With complete data sets mining data required as conditions of matter-element, it can find the gap between the quality of data, and promote the development transform of matter-element extension set domain with affair-element "data mining consulting", introducing measures to drive the data quality improvement through the application consulting of data mining, so as to solve the contradiction between data quality, so that the companies with low data quality can also implement data mining projects to improve decision-making information level.

Data mining application status and problems

In recent years, data mining has been more widely used in biology, finance, insurance, retail and other industries, and becomes the key technology of the information age. But in the process of data mining project negotiations and implementation, it often encounters a variety of contradictions. For example, there are contradictions between the quality of the data itself and the correctness of mining conclusions, between mass data and fast mining, the infiniteness of influence factors and limitedness of selected mining properties and so on. Among them, the data quality problem has become an important factor impacting the application of data mining, and the existence of incorrect or incomplete, redundant and sparse data make the reliability of final conclusion of data mining drop. Thus, the conclusions obtained by data mining experts from the enterprise with poor data quality can not be done data mining. Even if it does conclusions, the accuracy will be very low, can not be applied in commerce; In view of enterprises, the lack of effective measures can not make the data accurate, as well as can not effectively verify data accuracy, eventually resulting in long communication time for data mining projects. The corporate enthusiasm, from high to low, might make data mining project not come to an agreement. Due to the contradictions between the quality of data and the correctness of mining conclusions, the current business practices are usually to formulate standards for data quality inspection, to introduce of data quality management platform, to increase the intensity of punishment for data errors and purchase new technology and software, and these measures can not well settle the data quality problem, then enterprise information quality is still not high. Data mining experts put the focus on research data processing, cleaning techniques, algorithms, or mining algorithms research on low-quality data, but the effect is not ideal. Planning to set about the basic ideas, tools and methods of extension engineering researches, this paper will make formal description for the incompatibility issues, establish the conditions and purpose matter-element, regard the enterprise data as matter-element extension set, and make the analysis from the point of extension set change, trying to solve the conflict between poor data accuracy and high reliability required by data mining conclusion through extension transformation.

Extension analysis on data quality issues

Extenics sets the ordered triple R= (N, c, v) composed of object N, characteristic name c and v amount of N about c as the basic element of description for object N, called one-dimensional matter-element. Wherein an ordered two-tuples M = (c, v) made up of c and v represents a characteristic of the object N. According to the dynamic principle of Extenics, any matter-element is a function of the parameter t, namely R (t) = (N (t), c, v (t)), in which parameter t may be time, space or other parameters. Data used for mining is a dynamic multidimensional matter-element with time, space and information management degrees. From the time perspective, the initial software system of information is manipulation-oriented, and gives priority to improve work efficiency, and its content is incomplete, patches are more, then design implementation lacks unified planning; from the view of space, in the enterprise subsystems are relatively independent, and data is scattered. Various business norms and the basic data encoding lead to the diversity of expression ways; from the management perspective, there are problems such as data inconsistencies, incomplete data, duplicate data, data ambiguity and even the conflicts and other issues, but the management means and tools are short to find these problems. Furthermore, that the design did not provide a reasonable and effective way to update and maintain the data, lack of data quality supervision and management measures are also one of the reasons.

The most fundamental reason of poor data quality is that the data has not been effectively analyzed and applied by corporate executives.

The reason is that information systems started from business sector limited by information integration technology, and the data of each department formed information isolated island, data integrity and consistency can not be guaranteed. This kind of inaccurate, incomplete and isolated data is not conducive to subject-oriented analysis, not to mention the data mining, and decision support can not be effectively carried out. Therefore, the concerns of business leaders have reduced, and the driving force of accurate data has decreased, so that the data is less accurate, creating a vicious cycle. The application scheme of enterprise data mining based on extension transform involves three multidimensional matter-elements and a multidimensional affair-element, and raw data sets used for data mining can be expressed by multidimensional matter-element R; the quality should meet the requirements, and data sets available for effective mining can be showed by multidimensional matter-element k. The general data mining processes firstly transformed R1 into R2 through data cleansing, formatting, etc. to, and then adopted see5, support vector machines, McLP and other data mining software tools to achieve mining transformation, so as to the mining conclusion R3. Since data quality of companies is not high, data cleaning process often takes up a lot of manpower, material and time. Besides, data cleaning methods often take temporary solution and did not effect a permanent cure. Even though the laborious cleaning of the existing data, data collected for subsequent mining has been polluted by new and inaccurate data from the information system, which must be re-cleaned to mine and analyzed. The uncertainty of cleaning effect increases the risk of data mining projects. In the original data set for enterprises data mining can set up an extension set. The general definition of extension set is: set u as domain, k as a mapping to the real domain I from u, T=(TU，Tk，Tu) as a given conversion, A(t)={ (u，y，y’)| u∈TuU，y=K(u)∈I，y’=Tkk(Tu u)∈I，} is called a extension set on the domain TuU, y=k(u) is the correlation function of A(t), y’=TkK(Tu u) is extension function of A(r). In which TU, Tk, Tu are respectively the transformation for domain U, correlation function K (u) and element u. From the perspective of data mining, data R2 that can be effectively mined is a collection of complete, consistent and correct data associated with mining goal. An important task in data cleaning is to make data quality reach the requirements of effective mining through dirty data cleaning.

Solutions about domain transformation of discourse. To make displacement transformation for domain of discourse, you can choose the quality of the other data meeting the requirements of data mining to excavate, while changing mining goals; to make additions and deletions transformation for domain, you can increase data sets with better quality in order to reduce inaccuracy rate of the overall data set, or removing a little data with the poor quality and carrying on data mining with a subset after cleaning, are commonly used in data cleaning method, and the drawback is the large amount of cleaning work, easy to wash off a little of valuable information.

Solutions about correlation maxim transformation of discourse. Data set of enterprise for data mining data is invariant in itself, and that is the degree of correlation remains constant to convert the standards for judging the data quality, so that the data quality that does not conform to the requirements of general data mining software reaches the mining demand with the help of new software after conversion. For example, we can study and construct a data mining system of a low data quality, realizing the data mining algorithms which can tolerate low quality data.

Application examples

Since its establishment nine years ago, registered users and ordinary visitors of a website have gained rapid growth. Website content has become increasingly diverse, types of products are more and more, and all business units in the company have accumulated an increasing number of data, the value of which needs urgent analysis and excavation, providing decision support for the company's future development. In order to grasp the characteristics and the real needs of customers as soon as possible and develop products that meet customer needs, the company cooperates with data mining team from Chinese Academy of Sciences, and make in-depth analysis for with operational and customer data of the site with the aid of the extension theory and rich experience in data mining, then put forward the data mining programs implemented in phases with data mining consulting promoting the increase of data quality. Specific implementation steps are as follows:

1) Have a good understand of the overall data situation, and propose VIP mailboxes with relatively accuracy data as the analysis theme. Data research found that the information of mailbox user registration is less available, and valuable information exists in the log files, which has very short retention time and lacks effective and associated fields with database information. The program suggested to increase the information needed to collect data mining, and extend log file retention period, while converse the log formats (namely matter-element, "DM influence").

2) The enterprise should implement the improvement program data quality in line with recommendation.

3) After two months, re-examine the data and extract part of the data sample to analyze, then come up with optimal scheme of complaint information processing (namely matter-element, "DM influence"). Through three circulations, data can achieve completeness and accuracy of excavated requirements.

4) Data mining test. At present, decision trees, support vector machines, multi-objective linear programming and other methods have been adopted for test excavation, drawing some preliminary conclusions. Practice has proved that with Extenics as a guide, all data of enterprise can be considered as extension data collections by extension transformation. It found that data mining can be done in different ways of implementation no matter the level of data quality is high or low. Especially with matter-element introduction into extension set theory as a conversion means, the contradictory problems that companies of the low quality data can not carry on data mining will be worked out, making data mining expand the applications range to general business with low-rise data quality, and settling data quality problems of enterprise information fundamentally. This paper only attempted to solve practical problems by means of Extenics. How to make quantitative description for data sets quality applied to data mining with the help of correlation functions, will be the focus of the next study. In addition, in view of the low quality of data mining, there are other possible extension transformation programs, and we hope that more experts could take advantages of Extenics and other tools for researches.

References

[1] HAN　J，MICHEUNE　K．Data Mining：Concepts and Techniques[M]．[s．1]：Morgan Kaufmann，2006．

[2]KARGUPTA H，PARK B H，PITTIE s，et al．Contributed articles　on online，　interactive，and anytime data mining：　monitoring the stock market from a PDA[J].ACM SIGKDD Explorations Newsletter,　2002，3(2)：37—46．

[3] CHEN Yongqiang, HU Leifang. Study on data mining application in CRM system based on insurance trade[A]. Proceedings of the 7th　International Conference on Electronic Conference ICEC 05[C].[s.l.]:CM Press,2005.839-84