TELKOMNIKA, Vol.14, No.3A, September 2016, pp. 42~46
ISSN: 1693-6930, accredited A by DIKTI, Decree No: 58/DIKTI/Kep/2013
DOI: 10.12928/TELKOMNIKA.v14i3A.4383

Analysis on a New Data Mining Algorithm of the Statistics Work

Ding Zhi*, Ha Yan, Li Yan
Department of Computer Science and Technology, Bengbu University, Bengbu City, 233030, P. R. China
*Corresponding author, e-mail: dingzhibb@163.com

Abstract

The data mining is a massive, incomplete and random data. People don't know how to extract the data in advance because of the large amount of data of the network in the database, and the heterogeneous and dynamic characteristic, which makes the extraction process difficult. Data mining needs effective calculation. Cloud computing, as a business computing model, can assign a large number of calculate tasks. Based on users' need, it can allocate the computing, storage and application ability, and can solve massive computing problems effectively in data mining. According to the sequence of item location and the growth algorithm of SPM - LIPI mining, the algorithm discussed in this paper can improve the traditional algorithm, and form the same item at the end of the sequence position information table. It can also avoid repeated scans of the database mining through trimming by the database.

Keywords*: Statistical information; Data mining; Algorithm*

1. Introduction

Due to the rapid development of current information, especially the development of the popularization of large area at a high speed, the speed of updating information and producing information has been accelerated dramatically. The analyzing and selecting worthless information into a valuable intelligence has become an important means to improve the competitiveness of a country. Statistics department is very important in the whole system of government. However, modern developed countries always run without statistics department. Data statistics department is the basic part of the government for its macro decision-making. In the information technology revolution, people's life style and habits are changed, the development of information technology, the high speed of data processing and the transmission technology all coming into reality, which greatly shorten the delivery time of information, and improve the efficiency of decision-making. Therefore, making use of information technology will become increasingly modified. It will bring more reasonable statistics information, and perfect the statistical work process. Through its' maximum effect, the coordinated development of various industry overall must be significantly improved [1-5].

2. Research Method

2.1. An Analysis of the Data Mining Algorithm

The main effect of data mining algorithm is to extract the required information from a large amount of data, including the structural data, semi-structured, and unstructured data sources, such as audio, video, data, for data algorithm. This algorithm must have model, and first search algorithm. Currently the common data mining algorithms are mainly the decision tree method, bionic global optimization of genetic algorithm and neural network, statistical analysis and row exclusive counterexample method, etc. In order to improve the effect of data mining process effectively, a detailed research should be applied to the cloud computing method. In this way. More effective implicit knowledge in the mass data information can be discovered in order to improve the effect of application of information data [6-10].

Association rules found a relationship between things and other transactions or interdependence. Assuming that $I=\{i_1,i_2,\ldots i_m\}$ is a collection and the related data task D is a

Received March 8, 2016; Revised July 20, 2016; Accepted August 2, 2016

collection of database transactions, in which each transaction T is a collection, and making $T \subseteq I$. Every transaction has an identifier TD. Assuming that A is a set of items, $A \subseteq T$. Association rules are the containing type of $A \Longrightarrow B$, among them, $A \subseteq I$, $B \subseteq I$ and $A \cap B = \Phi$. The rules $A \Longrightarrow B$ is in the transaction which sets up with support S, S is the percentage for the transaction contains $A \cup B$ in the D.

2.2. The Goal of Data Mining and Tools

It is well-known that the data mining's fundamental purpose is to find the particular useful data structure of larger database based on its definitions. According to western scholars (D.J.H and), through the specific algorithms and data mining, the database model and the model structure can be excavated [11-13].

At present, many statisticians and econometricians, especially the researchers of the national statistics agency, their main work is to model and analyze according to the data model. They think it is necessary to do the data analysis and modeling. For them, model is like a summary, which can give a more comprehensive description of the variables and the relationships between variables. These two variables are the budget allowed to use the model results, which can help to understand the phenomenon of producing data. Under this circumstance, the linear models with the simultaneous equation can be used by these two variables in real analysis in widespread application. As a matter of fact, models are often used according to a specific and simplified theory (especially economic theory). But if using the algorithm of data mining, it will generate the stronger feasibility. Thus, data mining is using the data from different data mining techniques in the same concentration of different structure comparisons. Therefore the best way of predicting can be calculated. Some people think that as long as it is utilizing with the method of decision tree or neural network method and genetic algorithm, this can be called as data mining. This idea is very incomplete, since the data mining algorithms can provide a model based on widely related variables.

2.3. The Principle of LIPI Data Mining Algorithm

The LIPI algorithm through scanning data sets of frequency, then finding relevant data and finishing dig, its principle as follows:

Assuming that $I=\{i_1,i_2,\ldots i_n\}$ is a collection, which composed of different characteristics, the characteristics of each item as a constituted set of items. And the item set is not an empty set, but is a subset of the set of I, which can be expressed as$(x_1x_2\ldots x_m)$, every x_k is a term.

Table 1. Sequence database

S_{id}	Sequence
S_1	*ACAABC*
S_2	*ACABCB*
S_3	*ACABC*
S_4	*ABCAB*

Table 2. CA sequence database

S_{id}	Sequence
S_1	*ABC*
S_2	*BCB*
S_3	*BC*
S_4	*B*

The sample variance and sample proportion of variance have established the following relationships:

$$S^{*2} = \frac{S^2}{\mu^2} \tag{1}$$

Proof: by definition 4

$$S^{*2} = \frac{1}{n-1}\sum_{i=1}^{n}(\frac{\alpha_i - \overline{\mu}}{\overline{\mu}^2})^2$$
$$= \frac{1}{(n-1)\overline{\mu}^2}\sum_{i=1}^{n}(\alpha_i - \overline{\mu})^2$$
$$= \frac{1}{\overline{\mu}^2}S^2 \tag{2}$$

The algorithm is to propose the original data processing for many times, then use the effective information contained in the original data.

The study of the data mining algorithm has great significance in improving the effect of data processing. User data information needs to be extracted in the huge amounts of data, in order to promote the development of various fields. Cloud computing is a relatively new computing mode, its application in data mining also needs to do some further researches on the existing basis, continuously improving its application efficiency and improving data information of the application efficiency.

3. Results and Analysis

3.1. A Variety of Data Format, the Consistency and Create Data Redundancy Easily

The data mainly displays in some aspects, such as by the department of statistics data format, which does not have a uniform requirements. When using VF or SQL or some using TEXT or a variety of other formats, each system has general check, data format, and compatibility.

3.2. Exiting a Variety of Data Sources, Which are Scattered and Lack of Unified Management

Complicated statistical data have originated from the enterprises and institutions' direct submitting, or the result of the whole system of internal between different departments. For the lack of effective statistical data storage and management of professional means, it leads to the deep processing of statistical data.

3.3. The Lack of Abundant Data Resources in the Development and Utilization

It is known that the statistics business involves all aspects of society including index the great amount of data. Although there are abundant data resources, yet the lack of professional analysis of statistical data at a deeper level and lack of refining and mining tools, leads to the disunion between a large number of accurate data and resource usage and people's growing demand of statistical information.

3.4. The Statistical Data

The decision of government and enterprises and the current statistical work has been simply filled in form, and then submitted, but it is lack of effective means for the subsequent development. The data mining of data warehouse technology can solve the above mentioned problems effectively mainly because it has the following obvious advantages:

(1) Based on data warehouse algorithm, the data time-consuming preconditioning in data mining can be solved. Through the establishment of data warehouse, it can avoid data extraction, cleaning, conversion and loading process every time.

(2) Another feature of the data warehouse is to store data through subject organization, which provides convenience for data mining to choose the appropriate data source. According

to different areas, data warehouse is divided into the national economic statistics, social statistics analysis, and enterprise survey, etc. The national economy includes consumption statistics, labor statistics, people's living standard, and statistics, etc.

(3) The data is collected by statistics departments at all levels, all existing in different types of database such as EXCEL, Fox Pro, etc. Because historical data cannot be stored in the database, many knowledge can't be excavated in the mining database, such as forecasting and application; on the other hand, data warehouse storage management can get the data from the PLTP system and the history of offline business data and the external data sources of heterogeneous distributed, thus it's good for the heterogeneous data and source data to summarize in order to finish the more efficient usage. Except for the requirements of data mining and data warehouse environment, data mining needs to be based on data cube environment and data warehouse technology and hence meet the demands of data mining technology. Therefore, it is necessary for data mining and data warehouse to work together. On the one hand, data mining technology has become a very important application of data warehouse and a relatively independent tool; On the other hand, the data mining technology is an important step of catering the process of data mining, and it improves the efficiency of data mining and ability, and ensures that data mining have the extensiveness and completeness of data source.

4. Conclusions

Data mining is an emerging discipline of a high speed development. And data mining can be considered as a branch of exploratory statistics, it tries to find out the expected useful models and patterns. From the perspective of the problems of data quality, it is necessary to have a cautious attitude. However, the statisticians are familiar with risk and uncertainty, so they should decide the best candidate to test. From knowledge level, it is also likely to make errors, but still, it is very efficient. Therefore, human experience and intervention are necessary in the application of it. It is obvious that the new field of data mining has a well function in official statistics. Considering the yield of this new field and the large database, it is suggested that the statistics institutions should be redefined.

References

[1] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten. The WEKA data mining software. *ACM SIGKDD Explorations Newsletter*. 2011; (1): 203-205.
[2] Bernard J. Jansen. Search log analysis what it is, what's been done, how to do it. *Library and Information Science Research*. 2012; (3): 165-166.
[3] Craig Silverstein, Hannes Marais, Monika Henzinger, Michael Moricz. Analysis of a very large web search engine query log. *ACM SIGIR Forum*. 2011; (1): 834-835.
[4] Abdar M, Kalhori SRN, Sutikno T, Subroto IMI, Arji G. Comparing Performance of Data Mining Algorithms in Prediction Heart Diseases*. International Journal of Electrical and Computer Engineering (IJECE)*. 2015; 5(6): 1569-1576.
[5] Teevan J, Adar E, Jones R, Potts M. History repeats itself repeat queries in Yahoo's logs. *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. 2012; (08): 251-253.
[6] Zhang Z, Nasraoui O. Mining search engine query logs for query recommendation. *Proceedings of the 15th international conference on World Wide Web*. 2013; (08): 621-622.
[7] Dadkhah M, Sutikno T. Phishing or hijacking? Forgers hijacked DU journal by copying content of another authenticate journal*. Indonesian Journal of Electrical Engineering and Informatics (IJEEI).* 2015; 3(3): 119-120.
[8] Xu JX, Croft WB. Query expansion using local and global document analysis. *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* 2010; (12): 621-622.
[9] Shereen H. Ali, Ali I. El Desouky, Ahmed I. Saleh. A New Profile Learning Model for Recommendation System based on Machine Learning Technique. *Indonesian Journal of Electrical Engineering and Informatics*. 2016; 4(1): 81-92.
[10] Doug Beeferman, Adam Berger. Agglomerative clustering of a search engine query log. *Proc of the 6th ACM SIGKDD In-ternational Conference on Knowledge Discovery and Data Mining*. 2012; (02): 145-146.

[11] M R Narasinga Rao, Deepthi Gurram, Sai Mahathi Vadde, Sathish Tallam, N. Sai Chand, L. Kiran. A Predictive Model for Mining Opinions of an Educational Database Using Neural Networks. *International Journal of Electrical and Computer Engineering (IJECE).* 2015; 5(5): 1158-1163.
[12] Thorsten Joachims. Optimizing Search Engines using Clickthrough Data. *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD2002*. 2012; (06): 826-827.
[13] Hosseini, Mehdi, Hassan Abolhassani. Mining Search Engine Query Log for Evaluating Content and Structure of a Web Site. *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*. 2013; (05): 568-569.