International Journal of Database Theory and Application
Vol.9, No.6 (2016), pp.119-130
http://dx.doi.org/10.14257/ijdta.2016.9.6.12

Research of Association Rules Algorithm in Data Mining

Yingmei Xu

Department of Computer and Information Technology, Shangqiu Normal College
Henan Shangqiu, 476000, China
Qxuying2003@163.com

Abstract

With the rapid development of information technology, the data is becoming more and more large, the data mining technology has been applied to many fields in the life. In view of the existing data mining algorithms complexity is too high, the execution time too long, the article applies association rules in data mining technology, analysis summarized the existing advantages and disadvantages of the classic data mining association rules algorithm, is proposed for mining maximum frequent itemsets MFP - Miner algorithm. At last, by experimental simulation result shows that the proposed algorithm performance superiority, its complexity and execution time significantly shortened, the practical guiding significance.

***Keywords**: data mining; association rules; maximum frequent itemsets; MFP – Miner*

1. Introduction

With the rapid development of network and computer technology, more and more governments, organizations and scientific research units and education institutions to realize the information of digital processing. A large database [1], especially the data warehouse has been widely applied in product sales, computer science, business management and information services, and other fields, the rapid growth of the data to the database are also put forward higher request. Database management, storage and analysis needs to be further strengthen: on the one hand, the surge of hidden in the data is very important information, more people hope to have possession of information into the analysis of the high-level analysis, so that the data reasonable use; Facing the huge data, on the other hand, people need to develop new tools, so that the data collected by the automated processing and translated into useful knowledge and information. Although the current database system can achieve the function of data input, query and statistics, but also cannot be important information for huge amounts of data mining, which leads to the characteristic of "rich data, the lack of information".

The generation of data mining (DM) technology [2] was based on the above requirements. It not only can help people especially data warehouse data from a database to extract useful rules, knowledge or higher level of information, but also can help people from different levels to analyze them, and thus can be more effective use of the data warehouse or data in the database; It can not only describe the development of the data, but also can predict the trend of the development of the data. Therefore, data mining is becoming a hot research field of more and more attention. There are many research institutions and universities are engaged in work in the field of data mining, and has obtained the certain scientific research achievements.

As a kind of expression way of information, data can be considered as the main carrier of Internet communication, the progress of information technology at the same time, along with the explosive data. Since the birth of the Internet in various fields, especially in the increase of the amount of data in the field of information, the concept of big data. There is no doubt that information is becoming to promote human power is crucial for the development of social science and natural science. Along with the rapid development of information technology, human society has entered the era of big data.

Big data era, whether it is produced by the stock exchange trading information or a large number of log information generated by the Internet server, they cannot be directly used to human and understanding, the problem is in front of people, how to extract useful knowledge for human huge amounts of data, make more and more information known by people, the data mining technology.

Data mining technology has been applied to many aspects of the real life, can through the following analysis. Global retail giant Wal-Mart company in the consumer shopping behavior data analysis found that male customers in the supermarket to buy baby diapers, most of the time will be dropped in to buy a few bottles of beer, the analysis of data, a Wal-Mar supermarkets will diapers and beer on a piece of promotion methods, the results are surprising, whether diapers or beer, sales of both have improved significantly.

Examples above only from the Angle of the business of mining association rules [3] in data mining are given, in addition to these, the data mining technology in many fields such as telecommunications, finance, health, space give play to the role of with amazing, for example, the target keyword search using the search engines such as baidu, the server will be based on keyword matching in the server out of the millions of web information display to the user, but we want to use or can be we use the web information is rarely, if can use the classification method in data mining, people want to use more accurate access to web information. In addition, in gene technology, human DNA technology is also being a little crack, but the large DNA technology information and the relationship between some inextricably human diseases cannot be completely used by scientists. Through data mining techniques or by large populations of DNA information and joint disease information, can get all kinds of diseases and the relation between human cognition has not yet been DNA, help people to understand the cause of the disease, so as to make the effect a radical cure the disease become a reality.

Generally refers to the data mining from large amounts of data through relevant algorithm search out the hidden in the process of the information of them. Strictly speaking, this definition has three conditions: 1, the data mining found knowledge must be meaningful;2, the knowledge of data mining may not be widely received;3, the data source is not false and should contain noise. Data mining technique involves multiple areas, including the theory of artificial intelligence in automation technology, computer technology, database theory. In general, the analysis method of data mining mainly include clustering data mining method, method of classification of data mining, outlier data mining method and association rule data mining method, *etc.* For a variety of business issues, and using the corresponding analysis method can obtain more accurate conclusions.

In fact, in the field of artificial intelligence [4], data mining is being replaced by another term, namely knowledge discovery. The term more recognition in the field of Database data mining, so also known as Knowledge Discovery of Database, the data mining in English is called Knowledge Discovery in Database (KDD).In addition, one of the controversial idea that data mining is an important steps of the knowledge discovery, but not all, of its knowledge discovery, including data preprocessing, data integration, data cleaning, data distortion), data mining, knowledge discovery (including model evaluation, knowledge representation, *etc.*).In this paper, the research of data mining are thought to be an important step in the process of knowledge discovery, the special data mining.

2. Related Works

Data mining technology originated in the 1980 s database fields, in the later 10 years, obtained the rapid development of data mining technology, the 21st century, with the further development of information technology, data mining technology has got a very wide range of applications. Nowadays, with the rapid development of Internet technology, data mining technology is more and more shows the vitality of irreplaceable and ubiquitous influence.

In 1989, for the first time, KDD this noun in the 11th meeting of artificial intelligence; In 1995, the first KDD and Data mining, international academic conference on. Many countries in the world after 10 years of development, many excellent data mining software has been developed, and applied in many fields. In web data mining areas, for example, University of Minnersota developed web mime system; in the field of parallel data mining, the company developed the Hadoop open source software and graphs system; In the field of general data mining, some researchers have developed in New Zealand weka software, *etc.*

The purpose of the data mining is not just limited to the following five aspects: clustering, classification of data mining, data mining deviation, forecast data mining and data mining association rules of data mining [5]. These five functions are not independent, in data mining work, they will influence each other. The main method of data mining is put forward by the European and American countries. These methods mainly include artificial neural network method, the genetic method, the Rough Set method, decision tree method, *etc.* The method of artificial neural network is mainly influenced by biological theory, is composed of multiple independent input/output unit connection, the entire network is a strategy or function expression. Genetic method is mainly influenced by genetics theory, the genetic method by part: cross, replication and mutation, called the problem solution of chromosome genetic method, through the intersection of chromosome, replication and mutation to form new chromosomes, new chromosome by natural selection, and some were selected, others to be eliminated, of course, the natural selection here refers to the adaptability to problems, and then be selected new life activities between chromosomes, continue, finally get a best chromosomes, known as the optimal solution of the problem of answer.

Relative to the European and American countries, China in the aspect of data mining is relatively backward, but the development speed is amazing. It was not until 1997 that the first article on the research of data mining to appear in the journal of the China, in spite of this, China's commercial companies, research institutes and universities in the aspect of data mining and other professional research institutions or has made many achievements. Especially in the 21st century, China about the research of data mining are also rapidly emerging, and quality is improved. Rely on the rapid development of Internet industry, some researchers put forward many important in data mining algorithm and the theory of research, the research achievements of Chinese and foreign advanced further narrowing the gap between research institutions.

For example, in the study of classification learning, based on the characteristics of multitasking method of study, tsinghua university's Zhang Changshui team put forward a new learning method called rMTFL, first of all, the matrix can be used to explain the different characteristics of the method and the correlation between multiple tasks, and USES the Group Lasso thoughts related tasks of feature space is calculated, according to the results to obtain isolated tasks; Of southeast university Zhang Minling and others by using bayesian network the correlation between the class nature, so many class the learning problem is regarded as a series of single class label change, which makes the algorithm in multiple data sets in beyond the performance of the existing methods of mining; Jin Xiaomin, tsinghua university, people put forward the cross-domain active learning method; At the university of Beijing Wang Jianyong discriminant model by studying the uncertainty data mining problem of uHARMONY algorithm is proposed, the algorithm can be time consuming features do not need to choose, and find the discriminant model directly from the database, which makes uHARMONY classic uncertainty than the SVM classification algorithm, algorithm has greatly improve the using effect; In the use of data mining related data analysis for web information, whether it is from the Internet user behavior and relationships involved in the user plane data mining, and then to regional web information analysis, or from a high level of the network clustering coefficient evaluation, Chinese universities and research institutions have a lot of paper has carried on the detailed discussion. In addition to the above examples of data mining theory research, China about data mining has made some achievements in practical application. For example, social service providers in WeChat, microblogging platform to launch friend recommendation system based on data mining technology; On taobao alibaba launched the personalized recommendation based on association rule data mining theory system; Many Banks are using user credit model and so on, are all the data mining technology in the domestic important achievements of performance, in addition, in the field of artificial intelligence, data mining technology is also in constant development progress.

Now, researchers in data mining technology research hotspot focused on the following aspects:

For language study, data mining can expect results is similar to the SQL database [6] language standardization of data mining language, unified the standardization of language is helpful to promote the continuous development of data mining technology;

Another research focus is the visual representation of knowledge discovery, knowledge discovery conclusion is often expressed in mathematics, also often is the process of the knowledge discovery process and modeling language expression, this is not conducive to the human-computer interaction in the process of knowledge discovery and a better understanding of the conclusion.

Of web data mining and data mining of the popular research direction and the rapid development of the Internet makes it has many advantages in instead of traditional industry, with more and more industry to join the army of the Internet, the Internet the accumulated amount of data about user are also growing, the web data mining mainly reflects in: one is the analysis of website performance, the second is to improve web design, 3 it is the understanding of user behavior, by logging on the web site are transactional, so must be on the target before data mining of the web log mining data sets including session identification, task identification, path supplement related preprocessing operations. In the pretreatment method of the web log study, the author through study session identification algorithm of the method of combination of two kinds of judgment based on time. On the edge of the residence time threshold access request for the secondary judgment, improve the accuracy of the session identification.

On unstructured data and data mining, such as the multimedia data, text data such as data mining. These data is not exist, can be preserved by the database, data format in order to deal with the data in the database structure, you need to innovation a new data mining algorithm and the theory of data mining, data mining technology can be expanded mining object.

In addition, some other hot research directions include of space physics information data mining and data mining and other high-tech information of biological information data mining.

This article first on the related concepts in data mining association rules are introduced, according to the different standards for data mining association rules are classified. Secondly, the article to the problem of maximum frequent itemsets mining are analyzed in detail, and summarizes the results of previous studies, this paper proposes a maximum frequent itemsets based on the structure of FP-Tree MFP-Miner data mining algorithm, and finally the superiority of the proposed algorithm is verified by simulation experiment.

3. Proposed Scheme

3.1. Brief Association Rules

Hypothesis $W=\{w_1,w_2,..,w_3\}$ is a set of m different projects. Q is a collection of all transactions, namely transaction database, each transaction T is a collection of projects, T contained in W, *i.e.* $T\subseteq W$, and each transaction can use unique identifier to identify.

Definition 1: set X is a collection of W project, shorthand for itemsets, if $X\subseteq T$, then called transaction T contains X.

Association rules can be expressed as:, $X\Rightarrow Y$ and $X\subset W, Y\subset W$ $X\cap Y=\varnothing$,.The rule $X\Rightarrow Y$ in Q is constrained by the confidence and support. Confidence for the strength of the said rules, support to indicate frequency rules. It an be described as follows:

$$\text{confidence}(X\Rightarrow Y)=\mathrm{P}(Y|X)$$

$$\text{su pport}(X\Rightarrow Y)=\mathrm{P}(X\cup Y)$$

Definition 2: when using the association rules in data mining, the user needs to set up confidence and support in advance threshold, which produces in the process of mining association rules that could satisfy the requirement of the two threshold, to such a degree of confidence [7] and support are usually referred to as the minimum confidence and minimum support. To meet the minimum confidence and minimum support requirements of strong association rules called associations.

In this article, in order to better describe their relationships, respectively shorthand for a and b respectively and the confidence and support degree, minimum confidence $minconf$ and shorthand for minimum support $minsup$, and their values range between 0% and 100%.Also Q contains can be expressed $|Q|$ as the number of transactions, the number can be expressed as $|X|$ contained in X.

Definition 3: in Q the frequency of the project set X, which Q contains the number of transactions T, X called in Q support of the count, shorthand for $count$.

Based on the above support the definition of number and support, you can get the support of a project set X number and degree of relationship is $count=b\times|Q|$, in addition it can correspond to minimum support, the support number is defined as the minimum support threshold, shorthand for $mincount$, the relationship between it and minimum support can write to $mincount=minsup\times|Q|$.

Defining 4: set to the project X, if n included in a project X, so X called itemsets n.For example, $X=\{x_1,x_2\}$ is a set of two.

Define 5: if the project collection X support for not less than the minimum support, so called X for frequent itemsets, which X meet the requirements of minimum support. If a project Π meet the requirement of minimum support, and Π called for the frequent project, all the frequent item set called frequent 1 - itemsets, remember to L_1; Meet the requirements of the minimum support k itemsets called k frequent itemsets, all the set of frequent k itemsets L_k.In some literature called the frequent itemsets is frequent patterns, in this article unified by frequent itemsets.

3.2 data mining of maximum frequent itemsets

This section mainly introduced the basic concept of mining maximum frequent itemsets, and put forward a new algorithm based on frequent pattern tree MFP - Miner, the performance of the MFP - Miner algorithm carries on the detailed analysis and comparison and test [8].

Set to the project $X\subseteq W$, if $suppor(X)\geq minsup$, for X any superset Y, if $suppor(Y)<minsup$, X called for the Maximum Frequent item sets, or Maximum

Frequent patterns, and the Set of all X known as the Maximum Frequent itemsets collection of shorthand for MFS (Maximum Frequent Set).

Next, the article will be to the MFP - Miner's basic idea and basic theory in detail.

For any frequent itemsets R_i, the FP - Tree all contains the path to the R_i, can get through the nodes with the same chain R_i.

Through the process of FP - Tree the structure, can know all of the same name in the Tree node can nodes together is the same name, so for each path contains the node R_i can be obtained by traverse the nodes with the same chain.

Theorem 1: in the FP - Tree, if a node count is less than the minimum support count, then the node prefix and the path of the nodes in the item set for frequent itemsets.

To prove: assume that node N for path P suffix, and $\text{N.node-count} \geq mincount$, according to the construction process of the FP - Tree, can know to prefix path $\mathrm{P}^{'}$ of any one node must have: $N^{'}\text{.node-count} \geq \text{N.node-count} \geq mincount$, so you can know that the count N of all the nodes P in the smallest, made up of all the nodes in the pattern of count must be greater than or equal to $mincount$, is the frequent patterns. Never put off till tomorrow what you can.

For example, suppose P=<a:4,b:3,c:3,d:2> is a FP - Tree path, if set the minimum support count to 3, then you can get in front of the project c and the project of patterns for frequent patterns.

Theorem 2: if a frequent project a condition model base generated FP - Tree contains only a single path P, and then all of the projects P and set must be frequent itemsets, and the $\mathrm{P} \cup a$ support of the node number is equal to the middle of the number of support T.

Proof: according to the condition of frequent pattern tree structure can be know, for a particular project a frequently, in the condition of frequent patterns for frequent node in the tree node. Because the tree contains only a single path and path of each node are frequent node, a theorem 1 can know that the nodes in the path and the mode of project a must be frequent patterns.

According to the theorem 1 and theorem 2 can be obtained, certain conditions by the maximum frequent itemsets generated FP - Tree of frequent itemsets.Can therefore be deduced MFP - Miner algorithm, a specific algorithm description can be summarized as follows:

MFP - Miner algorithm:

Input: in $minsup$ the structure of FP - Tree, minimum support $minsup$.

Output: transaction database Q to meet the requirements $minsup$ of the maximum frequent itemsets collection of MFS.

```
(1) MFS = NULL;
(2) call MFP - Max (T, MFS, null);
(3) return MFS;
    Procedure  MFP-Max(T,MFS,x) / * found T all the suffix for x in the maximum frequent itemsets, co-exist in the MFS * /
(1) if T contains only a single path P then the begin
(2)  m=$\{a_1Ua_2U...Ua_n \mid a_i \in P\}$Ux and m.support=an.support;
/ * remove P contains all the project, and the support of the leaf node $a_n$ degrees assigned to m * /
(3)   if m is not a project in MFS set a subset of then
(4) the MFS = MFS∪m;/ * put m in MFS and remove the MFS is set * / m a subset of the project
(5) end
(6)   the else begin
(7) for k=HTable.length to 1 do begin
```

```
(8)  m=HTable[k].item-name∪x 且 m.support= HTable[k].support；.
(9)  structure condition frequent patterns and FP - Tree $T_m$ ;
(10) if  $T_m \neq$NULL then / * in addition to the root node also contains a child node * /
(11) calls MFP - Max ( $T_m$ , MFS, m);
(12) else if m is not a project in MFS set a subset of then
(13) MFS = MFS ∪ m;/ * put m in MFS, delete the MFS is m a subset of the item at the
     same time item set * /
(14)  End
```

4. The Experimental Results and Analysis

In order to verify and analyze the MFP - Miner algorithm performance, this section will be carried out a series of experiments, and the MFP - Miner algorithm and Mafia algorithm is used in the comparison. Is the MFP - Miner algorithm and Mafia algorithm comparison, because Mafia algorithm is considered to be the most promising algorithm.

4.1 Experimental Environment

In this paper, by using the test machine for desktop PCS, for Windows 7.0 operating system, memory is 512 MB, and use c + + as the programming tool, has realized the Mafia in visual c + + 6.0 and MFP - Miner validation of the algorithm.

In order to be more accurate to test the performance of MFP - Miner algorithm in this paper, we chose two types of databases, namely synthetic database and real database, the related parameters in the database is presented in Table 1.

Table 1. The Test Database

Transactional database	The number	Average length of the transaction	The transaction number	A collection of projects under a given minsup's top number
Connect4	140	41	68552	30
Chess	75	36	3210	21
T10I4D100K	1100	15	10000	12
Mushroom	130	22	8200	24

Real database is the biggest characteristic of the distribution of the frequent item set is dense, even in larger minimum support also can produce more higher dimensional maximum frequent item set, and the synthetic database T10I4D100K relative to the real database, it is less frequent item set dimension, and the maximum frequent item set of such database are mostly concentrated in 2 d.

4.2. Test and Analysis

In this paper, the experimental method is adopted, the database is selected first, and then under the different minimum support environment, run the Mafia algorithm and MFP - Miner algorithm respectively, and finally under the same parameters configuration, separately tested two different execution time of the algorithm. Here, this article only gives the execution time of the algorithm itself, not including data input and output. Because different input and output method of execution efficiency difference is bigger, if you include input and output, then will influence the accuracy of the algorithm. Therefore, this article tests, the test of time, saves the input and output is only test the execution time of the algorithm itself.

(1) on the Connect4 database for testing

Relative to the Chess database, Connect4 database contains more projects, the average transaction length is longer, and total number of transactions is many times more than Chess database, it contains the maximum frequent itemsets average length of about 18, and it also than Chess length of average maximum frequent itemsets in a database. Can be seen from Figure 1, MFP - Miner algorithm execution efficiency is about twice the Mafia algorithm. In addition, MFP - Miner algorithm shows better performance, it shows that algorithm for more Dave a few more effective data mining maximum frequent itemsets.

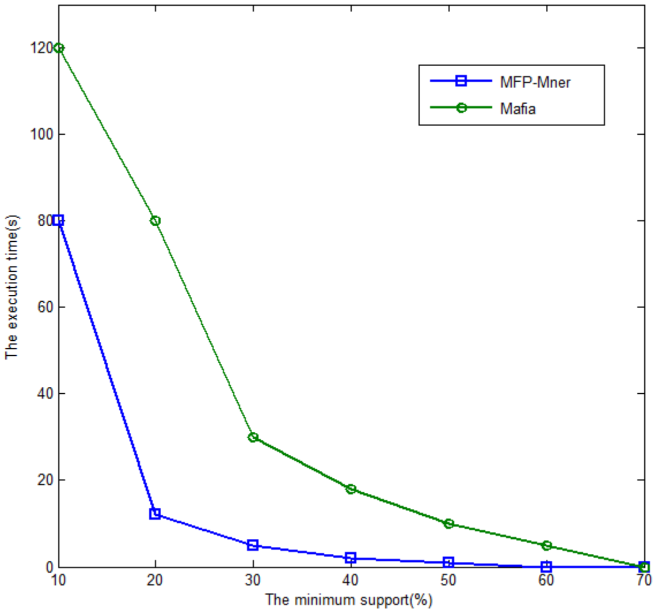

Figure 1. On Connect 4 Database for Testing

(2) On the Chess database for testing

Is the biggest characteristic of Chess database, the distribution of the maximum frequent itemsets is symmetrical, and the dimensions of the most frequent itemsets is lower, the average length of about 12.As can be seen from the Figure 2, when minimum support is more than 20%, MFP - Miner algorithm execution efficiency of 2 times faster than the Mafia algorithm. However, when the minimum support is less than 50%, MFP - Miner algorithm performance began to decline, and when the support is less than 30%, the performance degradation speed is faster. Can produce this kind of phenomenon, because MFP Miner algorithm using FP - Tree to compress the transactions in the database, and on this basis, due to take full advantage of the characteristics of the FP - Tree, in the process of data mining does not generate candidate itemsets, this will make it in the mining process has higher execution efficiency.

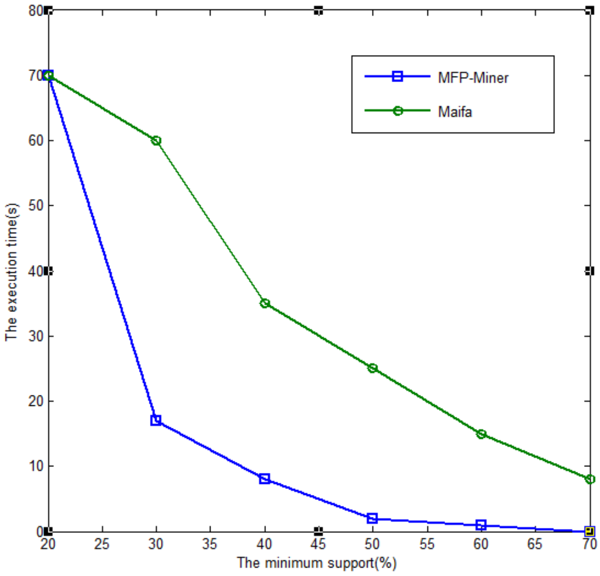

Figure 2. On Chess Database for Testing

(3) On the T10I4D100K database for testing

T10I4D100K database is characterized by its maximum frequent itemsets length is shorter, and focuses on the second dimension, but it contains a number of projects is the most. Through the Figure 3 as you can see, the minimum support between 0.02% and 1%, MFP - Miner algorithm in the change of the execution time is relatively stable, this means that the condition number of frequent pattern tree algorithm used in administrative overhead is relatively large.

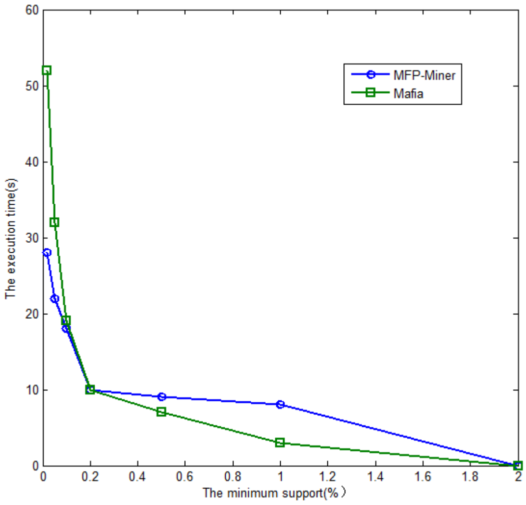

Figure 3. On T10I4D100K Database for Testing

(4) Test in Mushroom database

Mushroom database is the biggest characteristic of the distribution of the maximum frequent itemsets is concentrated, the length of each transaction is 22, and the vast majority of the length of the maximum frequent itemsets is 20, as a result, each maximum frequent item set has some programs exist in each of the transaction. As can be seen from the Figure 4, MFP - Miner algorithm on this database, its execution time is less, this suggests that the MFP - Mner algorithm for maximum frequent itemsets is longer and distribution of dense database has great advantage.

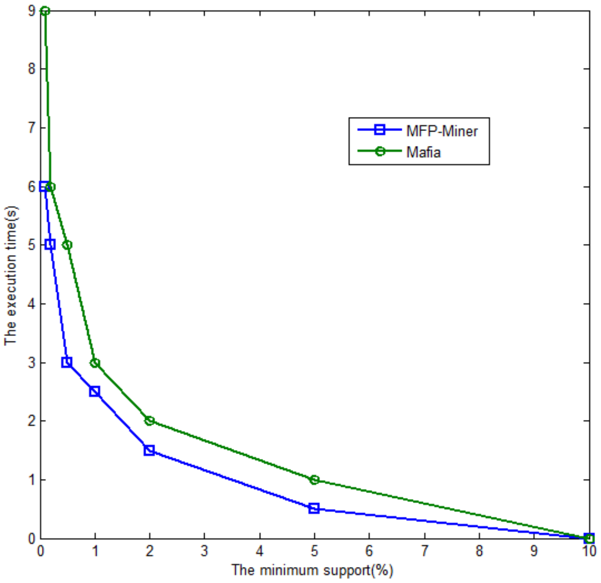

Figure 4. Test in Mushroom Database

5. Conclusion

Association rules reflect the relevant relationship between the project in a large amount of data set, in the transaction database, in the field of mining association rules of data mining is a very important research subject. This article first introduces the basic concepts of data mining and its application status quo, mainly studies the maximum frequent itemsets mining based on FP - Tree problem, put forward the MFP - Miner data mining algorithm, and through the experimental simulation, the performance of the proposed algorithm was analyzed, and pointed out the MFP - Miner algorithm in the superiority in the process of mining maximum frequent itemsets.

References

[1] M. B. Elshazly, R. Quispe and E. D. Michos, "Patient-Level Discordance in Population Percentiles of the Total Cholesterol to High-Density Lipoprotein Cholesterol Ratio in Comparison With Low-Density Lipoprotein Cholesterol and Non–High-Density Lipoprotein Cholesterol The Very Large Database of Lipids Study (VLDL-2B)", Circulation, vol. 132, no. 8, **(2015)**, pp. 667-676.

[2] P. Jiang and X. S. Liu, "Big data mining yields novel insights on cancer", Nature genetics, vol. 47, no. 2, **(2015)**, pp. 103-104.

[3] C. H. Weng, "Identifying association rules of specific later-marketed products", Applied Soft Computing, vol. 38, **(2016)**, pp. 518-529.

[4] A. Pannu, "Artificial Intelligence and its Application in Different Areas", Artificial Intelligence, vol. 4, no. 10, **(2015)**.

[5] A. Hamzaoui, Q. Malluhi and C. Clifton, "Association Rule Mining on Fragmented Database", Data Privacy Management, Autonomous Spontaneous Security, and Security Assurance. Springer International Publishing, **(2015)**, pp. 335-342.

[6] W. Lang, F. Bertsch and D. J. DeWitt, "Microsoft azure SQL database telemetry", Proceedings of the Sixth ACM Symposium on Cloud Computing. ACM, **(2015)**, pp. 189-194.

[7] J. Kubanek, L. H. Snyder and J. Hill, "Cortical alpha activity reflects the degree of confidence in committing to an action", Frontiers in Cellular Neuroscience, **(2015)**.

[8] N. West, J. Seong and E. Macdonald, "A randomized clinical study to measure the anti-erosion benefits of a stannous-containing sodium fluoride dentifrice", Journal of Indian Society of Periodontology, vol. 19, no. 2, **(2015)**, pp. 182.

Author

Yingmei Xu, Received B. Eng Degree in Henan Normal University in 1999 and 2003, and M. Eng Degree in China University of Mining and Technology in 2007-2010. She is currently researching on Internet of things, Data mining.