Australian Journal of Basic and Applied Sciences, 5(10): 536-539, 2011
ISSN 1991-8178

Development and Application of Parallel Distributed Data Mining Algorithm in Optimal Management of Databases for Organizations

Fatimah Masoudifar

Sama College of Azad-Shahr, Azad-Shahr, Iran.

Abstract: Regarding the exponential growth of information throughout the world, the companies are always engaged with a great deal of digital information growing. One of the most important challenges for data mining is finding the right and quick relation between data. Apriori algorithm is the most popular technique for discovering the repetitious patterns. Nevertheless, when we are using this technique, a database should be scanned numerously to calculate many sets of chosen items. Parallel and distributed calculations are among the effective strategies to accelerate the data mining process. In this paper, parallel distributed Apriori algorithm has been introduced as a solution for the problem. In this method, only one scan from database is needed. The procedure considers one factor from a collection of items. Therefore, production of the equilibrial work volume among the processors and reduction of the idle processor time is mentioned. In order to show the work of the proposed technique, it is compared with some other parallel algorithms of data mining. Finally using structural procedures, credibility and stability of the algorithm is studied.

Key words: Algorithm, data mining, parallel, distributed, organization.

INTRODUCTION

Data mining or exploring the knowledge in databases is a relatively new science which seems to be needed more and more with recent advancements of our country in IT, special attentions to the electronic government and the prevalence of using computerized systems in industry and creating huge databases by governmental and nongovernmental offices, banks and private sectors. Data mining deals with exploring the credible knowledge and information of databases. In other words, data analysis by machines to find useful, new and documentable patterns in large databases is called data mining. Data mining is also very common for small databases whose obtained results and patterns can be utilized for making strategic commercial decisions of small companies. The usage of data mining can be summarized in one phrase: "data mining provides the information you need for making intelligent plans about serious difficulties of your job" (Agawal *et al.*, 1993).

With fast development in information technology, companies have been working to digitalize all aspects of their activities aiming to improve efficiency and competition. A great deal of data has been produced by complete digitalization which is very important for extracting meaningful information from the distributed data. Data mining techniques have been developed recently for this purpose, meanwhile being categorized in different models such as categorized regression, time series, clustering, assembly, continuous and among others. Assembly regulations are being used in many practical programs, one critical step of which in data mining is exploring the repetitious patterns. This step needs time intervals in which patterns appear in the database. Regarding the methods to produce selective patterns, research can be categorized into production and examination of algorithms such as Apriori, Agrawall, Han, Prince Edward Iceland, Yin and Mao. Heretofore, an upward procedure was used which was the repetitious extension of a subject's subsets at one specific time. Although in many techniques proposed such as Apriori, it takes long time to find repetitious patterns whose database includes many interactions. Some researches on parallel and distributed techniques have accelerated the exploring procedures effectively. Charges of irregular and non-equilibrium calculation could be decreased dramatically due to the total operation. Charge equilibrium in processors for parallel and distributed exploration is very critical through the data mining processes. In this paper, the parallel distributed algorithm of Apriori has been presented for solving the problem whose objective is to reduce the scanning frequency of databases and also to establish equilibrium between the involved and calculated charges in the calculation node. In this method, database is scanned just once because the ultra-data is stored if the id was scrutinized before. This procedure is also used both in counting the set of items in order to improve the charge equilibrium and in reducing the idle time of the processor. The experimental results reveal that the time for implementing the proposed method is significantly less than some previous methods. Results also imply that the algorithm has led to lower scans and it can be distributed equally through the processors (Agawal and Srikant, 1994).

This paper continues as following: Section 2 studies some ideas briefly which are the base in forming the algorithms and their rules by reviewing the literature. Section 3 deals with describing the Apriori algorithm and operation test, it also explains the implementation route for the Apriori algorithm. Section 4 brings some case

Corresponding Author: Fatimah Masoudifar, Sama College of Azad-Shahr, Azad-Shahr, Iran.
E-mail: fatememasoudi@gmail.com

studies while sections 5 and 6 speak about the results and discussion, conclusions and proposals for further studies.

Literature Review:

The Apriori algorithm introduced by agrawall and his colleagues in 1994 is one of the most important algorithms available in exploring the patterns of data mining. Its main idea implies that the observations are based on a subset from repetitious collections. Apriori algorithm tests repetitious items by one item at one time and expands it. The final condition relates to the time when no other successful form is accessible. This algorithm can find the patterns once it repeats so they can be useful in reducing the search time and cost (Agrawal and Shafer, 1996).

Agrawall and Shaffer introduced the parallel algorithms based on count distribution and data distribution in 1996 in order to solve the problems of data mining. Their structure is appropriate for data mining but regarding to the huge volume of data and the low speed of search in these algorithms, researchers have tried to increase their search speed. Chaung et al invented a procedure of data mining based on fast distribution which led to lower data mining times. In addition, the quick parallel data mining process was prepared in 2002 by Chaung which played key role in improving the efficiency of data mining (Apte and Weiss, 1997).

Ye and Chiang presented a parallel distributed algorithm based on a tree structure which offered higher speeds of data communication in the data mining structure. Lee and coworkers have prepared an efficient repetitious data mining algorithm recently. It consists of a novel structure for data storing which has provided the required flexibility and speed. Studying the algorithms can reveal that each of the algorithms has succeeded in meeting the special requirements of data mining. Thus it is tried to provide an efficient pattern with parallel distribution and high speed of search in order to discover the pattern by one repetition. This is based on Apriori algorithm which will be discussed later in the next section.

Parallel Distributed Data Mining Algorithm of Apriori:

Running time for different processors in the algorithm presented by Chiang and Yi may have significant variations by time. In order to prevent long running times, unbalance in distribution of different charges in the database and frequent repetitions, Apriori algorithm has been presented here. First it is required to obtain the weight of distributed charges which can be done by using formulae (1).

$$\text{weight}(l_i) = \text{len}(\text{freq}_k) - i - 1 \quad (1)$$

Where, $\text{len}(\text{freq}_k)$ is total repetitions for the set of K-th item and Li is the amount of charge distributed. For calculation of the total weight we use formulae (2).

$$\text{TotalWeight} = \sum_{i=0}^{\text{len}(\text{freq}_k)-1} \text{weight}(l_i) \quad (2)$$

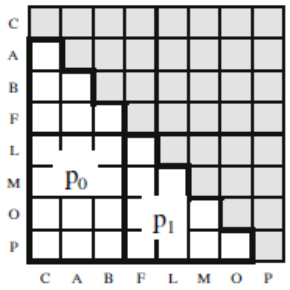
General scheme of the Apriori algorithm is as follows:
Consider that we show the set of data like this: DB-$\{T_0, T_1, \ldots, T_{a-1}\}$ and $T_i=\{i_0, i_1, \ldots. I_{m-1}\}$ then we can be able to explain the steps of algorithm:

Step1: Each processor reads data from the information database (DB).
Step2: Each processor scans DB and creates a collection of interaction items for any them.
Step3: Each processor counts the number of repeated sets.
Step4: The K parameter will take the value of 1.
Step5: The mother processor categorizes the repetitions and the first processor will get a set of the candidate items.
Step6: Each processor counts repetitions of the candidate item and sends the results to the mother processor. Let K be K+1.

In the following and based on the presented algorithm, we will provide a case study.

Case Study:

In order to study the presented algorithms, we have tried to run the mentioned pattern on a database in Alupan Company. Data results based on the algorithm is summarized in figures 1 to 3. In figure1, the number of distributed repetitions in 2 processors is defined.



Item	TID
C	1345
A	125
B	345
F	23
L	24
M	23
O	25
P	15

Fig. 1: Distributed repetitions in 2 processors.

In the above table, two processors of P_0 and P_1 can be seen. The number of networks is 28.
In the following it is tried to show the number of database scans in order to find one item in figure2.

TID	Items
1	ACDGP
2	AFLMO
3	BFCM
4	BCKSL
5	ACOPB

→

Item	TID
A	125
B	345
C	1345
D	1
F	23
G	1
K	4
L	24
M	23
O	25
P	15
S	4

Fig. 2: The scan of database in order to find one item for candidate.

The final algorithm based results based are depicted in fig. 3.

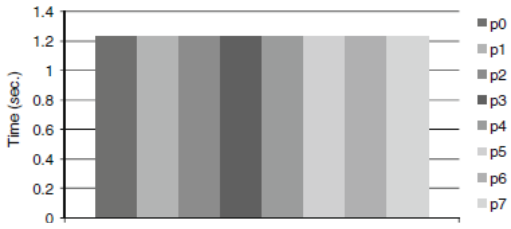


Fig. 3: Running times of algorithms in processors.

Next, for studying the model's stability it is tried to set interviews with 16 experts from universities and organizations. 85% of them approved the model. The remained 15% change their opinion by modifications done through 7 revisions in the apparent structure of the algorithm, so they accepted the model finally.

Discussion:

The combined algorithm presented was evaluated and it was shown that the introduced pattern with highest speed and lowest standard deviation was the best optimum pattern. The next patterns of Ye and Edma were chosen as the appropriate patterns having high speed and low standard deviation. The mentioned structure can be seen in fig.4.

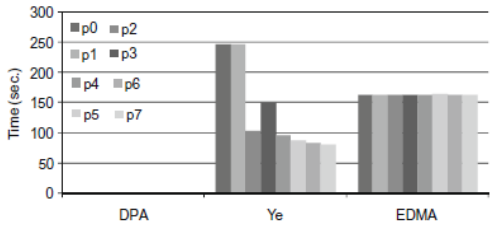


Fig. 4: Comparison between different algorithms.

The pattern was studied from the stability point of view and it was approved by academic and industrial experts of this field. The pattern is compatible to be implemented in both industrial and academic organizations. It is the first time that this type of optimization pattern and/or algorithm is introduced.

Conclusions:

In this paper, one parallel distributed data mining in optimal management of organizational databases is proposed whose algorithm structure is such that it leads to lower running times and repetitions of search during the data mining. The mentioned algorithm was studied in the database of Alupan Company and its obtained results were compared with results from the available algorithms. In addition to creditability, its stability was also scrutinized by structural procedures.

Further Studies:

One good proposal for optimizing the structure introduced, was using some other proper search structures based on neural networks.

REFERENCES

Agawal, R., T. Imilinski and A. Swami, 1993. Mining association rules between sets of items in large databases. In Proceedings of the 1993 ACM SIGMOD international conference on management of data, 22(2): 207-216).

Agrawal, R. and J.C. Shafer, 1996. Parallel mining of association rules. IEEE Transactions on Knowledge and Data Engineering, 8(6): 962-969.

Agrawal, R. and R. Srikant, 1994. Fast algorithms for mining association rules. In Proceedings of the 20th international conference on very large databases, pp: 487-499.

Apte, C. and S.M. Weiss, 1997. Data mining with decision trees and decision rules. Future Generation Computer Systems, 13(2-3): 197-210.

Bodon, F., 2003. A fast Apriori implementation. In Proceedings of the IEEE ICDM workshop on frequent itemset mining implementations.

Cheung, D.W., J. Han, V.T. Ng, A.W. Fu and Y. Fu, 1996. A fast distributed algorithm for mining association rules. In The fourth international conference on parallel and distributed information systems, pp: 31-42.

Cheung, D.W., S.D. Lee and Y. Xiao, 2002. Effect of data skewness and workload balance in parallel data mining. IEEE Transactions on Knowledge and Data Engineering, 14(3): 498-514.

Cheung, D.W., V.T. Ng and A.W. Fu, 1996. Efficient mining of association rules in distributed databases. IEEE Transactions on Knowledge and Data Engineering, 8(6): 911-922.

Einakian, S. and M. Ghanbari, 2006. Parallel implementation of association rules in data mining. In Proceedings of the 38th southeastern symposium on system theory, pp: 21-26.

Han, J., J. Pei, Y. Yin and R. Mao, 2004. Mining frequent patterns without candidate generation: A frequent-pattern tree approach. Journal of Data Mining and Knowledge Discovery, 8(1): 53-87.

IBM Almaden, 1994. I. Quest synthetic data generation code. <http://www.almaden.ibm.com/cs/quest/syndata.html>.

Parthasarathy, S., M.J. Zaki, M. Ogihara and W. Li, 2001. Parallel data mining for association rules on shared-memory systems. Knowledge and Information Systems, 3(1): 1-29.

Wu, J. and X.M. Li, 2008. An efficient association rule mining algorithm in distributed database. In International Workshop on Knowledge Discovery and Data mining (WKDD), pp: 108-113.

Ye, Y. and C.C. Chiang, 2006. A parallel apriori algorithm for frequent itemsets mining. In Proceedings of the fourth international conference on software engineering research, management and applications, pp: 87-94.

Zaki, M.J., M. Ogihara, S. Parthasarathy and W. Li, 1996. Parallel data mining for association rules on shared-memory multi-processors. In Proceedings of the 1996 ACM/IEEE conference on upercomputing (CDROM).

Zaki, M.J., S. Parthasarathy, M. Ogihara and W. Li, 1997. Parallel algorithms for discovery of association rules. Data Mining and Knowledge Discovery, 1(4): 343-373.