ASSOCIATION RULE MINING: A DATA PROFILING AND PROSPECTIVE APPROACH

Hemant Kumar Soni[1], Sanjiv Sharma[2], Pankaj K. Mishra[3]
[1]Asst. Prof., Dept. of Computer Science, ASET, Amity University, Gwalior (M.P)
[2]Asst. Prof., Dept. of Computer Science, MITS, Gwalior (M.P)
[3]Asst. Prof., Dept. of Applied Science, ASET, Amity University, Gwalior (M.P)
Email:soni_hemant@rediffmail.com[1]

Abstract

The Main objective of data mining is to find out the new, unknown and unpredictable information from huge database, which is useful and helps in decision making. There are number of techniques used in data mining to identify frequent pattern and mining rules includes clusters analysis, anomaly detection, association rule mining etc. In this paper we discuss the main concepts of association rule mining, their stages and industries demands of data mining. The pitfalls in the existing techniques of association rule mining and future direction is also present.

Keywords: Association Rule Mining, Frequent Pattern, Apriori, FP-Tree, Incremental data mining, support, confidence.

Introduction

Data Mining is the iterative and interactive process of discovering valid, novel, useful, and understandable and hidden patterns. Data Mining is used in extracting valuable information in large volumes of data using exploration and analysis. With an enormous amount of data stored in databases and data warehouses requires powerful tools for analysis and discovery of frequent patterns and association rules. In data mining, Association Rule Mining (ARM) is one of the important areas of research, and requires more attention to explore rigorously because it is an prominent part of Knowledge Discovery in Databases (KDD).

Application area of data mining is very vast, such as Remote Sensing, Geographical Information System, Cartography, environmental assessment & planning a name of few.

Association Rule Mining

Recently, researchers are applying the association rules to a wide variety of application domains such as Relational Databases, Data Warehouses, Transactional Databases, and Advanced Database Systems like Object-Relational, Spatial and Temporal, Time-Series, Multimedia, Text, Heterogeneous, Legacy, Distributed, and web data [1].

Since data generated day by day activities, the volume of data is increasing dramatically. Massive amount of data is available in the data warehouses. Therefore, mining association rules helps in many business decision making processes. Some examples are cross-marketing, Basket data analysis and promotion assortment etc. In the area of association rules mining, a lot of studies have been done. The association rules mining first introduced in [2] [3] [4].

For a given transaction database T, An association rule is an implication of the form $X \Rightarrow Y$, where $X \subset I$, $Y \subset I$, and $X \cap Y = \Phi$, i.e. X and Y are two non-empty and non-intersecting itemsets. The rule $X \Rightarrow Y$ holds in the transaction set D with confidence c if c % of transactions in T that contain X also contain Y.

A transaction T is said to support an item i_k, if i_k is present in T. T is said to support a subset of items $X \subseteq I$, if T support each item i_k in X. An itemset $X \subseteq I$ have a support s in D. It is denoted by s(X). If s% of transactions in D support X.

There is a user-defined minimum support threshold, which is a fraction, i.e., a number in [0, 1].

Support $(X \Rightarrow Y)$ = Support $(X \cup Y)$ / |D| ------ (1)

The confidence of rule X ⇒ Y is the fraction of transactions in D containing X that also contain Y. It indicates the strength of rule.

$(X \Rightarrow Y)$= Support $(X \cup Y)$ / Support (X) ------ (2)

Stages in Association Rule Mining

In [3], the problem of discovering association rules is decomposed into two stages: Discovering all frequent patterns represented by large itemsets in the database, and generating the association rules from those frequent itemsets. The second sub problem is a straightforward problem, and can be managed in polynomial time. On the other hand, the first task is difficult especially for large databases. The Apriori [3] is the first efficient algorithm for solving the association rule mining, and many of the forthcoming algorithms are based on this algorithm.

Confidence denotes the strength of implication and support indicates the frequencies of the occurring patterns in the rule. It is often desirable to pay attention to only that rule which may have reasonably large support. Such rules with high confidence and strong support are referred to as strong rules [2] [5]. The prime objective of mining association rules is to discover strong association rules in large databases.

Applications of Data Mining

As data mining matures, new and increasingly innovative applications for it emerge. Although a wide variety of data mining scenarios can be described. Applications of data mining are divided as follows:

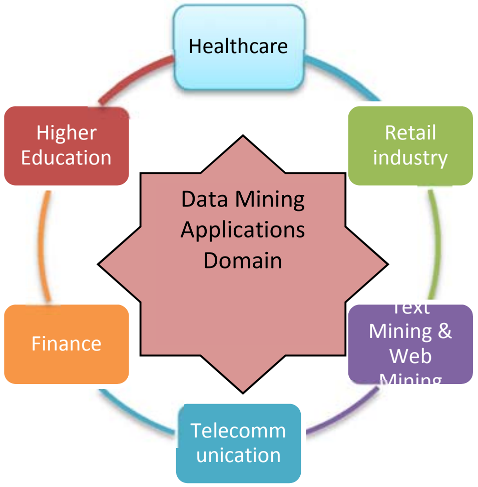

Figure 1 Data Mining Applications Domain

1.5 Industries demand of Data Mining

Data mining attract all kind of companies to provide valuable information that helps them to stay in the competition. It also help them in effective and efficient decision making and in future planning. Data mining is popular and successful in many areas in which data base marketing and credit card fraud detection is most popular. For example, the following areas of database marketing, in which data mining play an important role [6].

a. ***In Response Modeling*** : Based on previous history and other data like demographic, geographic and life style data, data mining predicts which prospects are likely to buy.
b. ***In Cross Selling*** : Based on the purchase pattern and frequently purchased items data mining helps in increases sales providing effective services to the existing customers.
c. ***In Customer retention***: Based on the customer buying habits and purchase patterns and analyzing the competitor's policies, data mining helps in making strategies for customer retention.
d. ***In segmentation and profiling*** : through classification and clustering, data mining helps in segmentation and profiling customers.

With the use of data base and information technology, data mining is valuable and useful in any industry or business sector. Some of the applications are given below [7] [8]:

- **Fraud detection**: Data mining helps in identifying fraudulent transactions.
- **Loan approval**: data mining techniques helps in establishing credit worthiness of a customer requesting a loan.
- **Investment analysis**: Based on the historical database, data mining predicting a portfolio's return on investment.
- **Portfolio trading**: Data mining support in trading a portfolio of financial instruments by maximizing returns and minimizing risks.
- **Marketing and sales data analysis**: Data Mining help in identifying potential customers; establishing the effectiveness of a sale campaign.
- **Manufacturing process analysis**: In manufacturing, data mining helps in

identifying the causes of manufacturing problems.

- **Scientific data analysis**: large scientific data can also analyze with the help of data mining techniques.

Incremental Data Mining

Transaction database will increase in volume with the time. Since the database updated and increases, association rule mined from old database requires to be revaluated. Database updation will change the support and confidence value, hence old association rule may turn out to be invalid and some new association rule may be relevant [9] [10]. Batch mining concept used by Apriori and FP-Tree mining Algorithm. Conservative batch mining algorithm like Apriori and FP-Tree algorithm resolves the incremental mining problem by re- processing of the entire new database, when new transactions are inserted.

Conclusion

Most of the researchers have considered association rule mining problems as single objective problem and validated on static database but it is a multi-objective problem because it uses measures like support count, comprehensibility and interestingness for mining the strong association rule [11] [12] [13] [14]. Since the database is being updated periodically due to daily business requirement. Incremental mining deals with generating association rules from updated database.

Most of the existing algorithms for incremental mining rescan the entire database again. Cost of scanning large database is high. The association rules generated on static database are not meaningful, effective and not appropriate for making business strategies and decisions. Hence, it requires to devise a new and efficient algorithms for incremental mining without rescanning of database. Therefore, there is a need to shift the paradigm form single objective to multi-objective association rule mining and also requires consideration of incremental data.

Data mining is a new and significant area of research, and soft computing tools itself are extremely appropriate to solve the problems. Soft computing characteristics include high robustness, parallel processing, self organizing adaptive, high degree of fault tolerance distributed storage etc are much suitable for data mining applications. It also obtain a greater attention in Artificial Neural Networks, which offer qualitative methods for business and economic systems.

Reference

1. Luo, D., Cao, L., Luo, C., Zhang, C., and Wang, W. “Towards business interestingness in actionable knowledge discovery”, IOS Press, Vol. 177, pp 101–111, 2008.

2. R. Agrawal ,T. Imielinski and A. Swami, “Mining Association Rules between sets of Items in large Databases”. In Proc. 1993 ACM-SIGMOD, Int. Conf. Management of Data (SIGMOD’93), pp 207-216 Washington ,DC, May 1993

3. R. Agarwal and R. Srikant, “Fast Algorithm for Mining Association Rules”. In Proc. 1994, Int. conf. Very Large Data Bases (VLD’94), pp 487-499, Santiago, Chile, Sept’94.

4. M. Houtsma and A. Swami. “Set-Oriented Mining of Association Rules”. Research Report RJ 9567, IBM Almaden Research Centre ,San Jose ,California, Oct.’93.

5. Piatetsky – Shapire (Editor),“Knowledge Discovery in Databases” ,AAAI/MIT Press ,1991

6. Chengqi Zhang Shichao Zhang, “Association Rule Mining - Models and Algorithms”, *Springer* , 2002

7. U. M. Fayyad and E. Simoudis, “Data mining and knowledge discovery”. In: *Proceedings of 1st International Conf. Prac. App. KDD& Data Mining*, pp 3-16, 1997.

8. G. Piatetsky - Shapiro and C. Matheus,“ Knowledge discovery workbench for exploring business databases”. *International Journal of Intelligent Systems*, 7, pp 675-686, 1992.

9. W. Cheung and O. R. Zaiane. “Incremental Mining of Frequent Patterns without Candidate Generation or Support Constraint”. Proceedings of the 7th International Database Engineering and Application Symposium, pp 111- 116, July 2003

10. B.Xu ,T.Yi, F Wu and Z Chen. “An incremental updating algorithm for mining association rules”, Journal of electronics, pp 403-407,2002.

11. P. Wakabi-Waiswa and V. Baryamureeba, "Mining High Quality Association Rules using Genetic Algorithms", In *Proceedings of the twenty second Midwest Artificial Intelligence and Cognitive Science Conference*, pp. 73-78, 2009.

12. M. Anandhavalli and S. Kumar Sudhanshu, A. Kumar and M.K. Ghose, “Optimized

Association Rule Mining Using Genetic Algorithm", *Advances in Information Mining*, vol. 1, issue 2, pp. 01-04, 2009.

13. Ghosh S., Biswas S., Sarkar Dand Sarkar P.P., "Mining Frequent Itemsets Using Genetic Algorithm", *International Journal of Artificial Intelligence & Applications*, Vol. 1, No. 4, pp. 133-143, 2010

14. W. Soto and A. Olaya-Benavides, "A Genetic Algorithm for Discovery of Association Rules." In *Computer Science Society (SCCC)*, pp. 289-293, 2011.