Application of Incremental Mining and Apriori Algorithm on Library Transactional Database

Gunjan Mehta
Bahra University,
Solan, HP, INDIA

Deepa Sharma
Bahra University
Solan, HP, INDIA

EktaChauhan
Bahra University
Solan, HP, INDIA

ABSTRACT

Data mining is used to extract hidden, predictive information from large databases, which can be used for predicting future trends and allowing businesses to make knowledge- driven decisions [1].In this paper we explain how Apriori algorithm can be applied on the university's library transactional database in order to find out the frequent book items and generate rules on these book items so as to predict the book borrowing behavior of the students. It then explains how incremental mining when incorporated by adding five more transactions to the original set of ten transactions changes the number of frequent item-sets and association rules generated by the algorithm.

General Terms

Data mining algorithms, library information system, frequent item-sets

Keywords

Apriori algorithm, associations rule mining, incremental data mining.

1. INTRODUCTION

Libraries are an integral part of any organization. They form the basis of every education system and therefore must be well organized so that it can be used and managed efficiently. The library management systems available today are designed to manage the books, journals, manuscripts etc. They keep track of the books issued or returned and tell about the availability of the books in the libraries. It is clear that the usage of these library management systems is very limited. These systems do not keep track of the operational or transactional details occurred in the library over a longer duration of time. It is obvious that the usage patterns of the library books and journals developed over the years are embedded in these transactions. If these patterns can be discovered then these can effectively influence the library operations, investment and procurement plans for the future growth of a library and so on [2]. Therefore there is a need for such a computer based library management system that makes the task of managing the libraries efficiently much easier and is also intelligent enough to extract hidden patterns from the transactions and support decision making. At the beginning of each adjustment in a library, librarians or library managers need to think about which kind of books are going to purchase or what kind of magazines have to be ordered in the coming renewal period of the year. By incorporating a decision support system it will help in making statistical analysis. It will offer librarians sufficient information to make right decisions. Reader's behavior can be derived and used to improve or modify the present library service. Thus huge benefits can be gained by incorporating a decision support system in the library. Such a system can only be implemented by applying data mining to the data warehouse of the library.

1.1 Data Mining

It is a process of extracting hidden and meaningful information in the form of patterns and relationships from large data warehouses that contain historic data. This information is used to predict future trends and support decisions makers. This information is gained by extracting one or more of the following relations.1) *Classes*. The data in the warehouse can be classified under predetermined groups.2) *Clusters*: this is a non-supervised classifying task in which the groups are logical but not pre-defined. 3) *Associations*: These are relationships that define inter-dependence among various variables or data items. 4) *Sequential patterns*: define sequence based relationships or trends.

1.2 Association Rules

Frequent item set mining and association rule induction [3] are powerful methods for the market basket analysis, which aims to find regularities in the shopping behavior of customers of supermarkets, mail-order companies, online shops etc. These rules in simplest form are like If-Then statements such as 'person buys bread then he will most probably also buy butter'. Such rules help in designing better super markets, policy plans, e-shopping website designs etc. First frequent item-sets in a data base are found which have support greater than minimum support value and then based on the confidence of occurrence of two or more frequent item-sets together association rules are generated that define relationship among such items.

Apriori is one of the most popular algorithms proposed by R. Agrawal and R. Srikant in 1994 for mining frequent item-sets and generating association rules [3]. It is designed to operate on a transactional database. For a given set of item-sets, the algorithm tries to find subsets of items which are common to at least a minimum number of K item-sets. Apriori uses a "bottom up" approach, where the frequent item-sets are extended one item at a time, this is called candidate generation. The algorithm terminates when no further successful extensions are found. The purpose of the Apriori Algorithm is to find associations among different item-sets. The Apriori process is as follows:

1. Find all item-sets that have minimum support (frequent item-sets, also called large item-sets).

2. Use frequent item-sets to generate rules.

1.3 Incremental Mining

Most mining algorithms which mine frequent item-sets from a large database efficiently work on a static database which has no scope for updation. However, all transactional databases are dynamic and are incremented regularly. For example, the library transaction database is growing daily since everyday books are borrowed or returned. Therefore, the database keeps on changing and is never static. The existing algorithms are not suitable for such databases since the association rules generated for the old database will no longer be valid for the updated database. Also to generate the rules for the entire database every time it is updated is tedious. Algorithms that make use of the results of the previous mining exercise for re-mining are referred to as incremental mining algorithms.

2. LITERATURE REVIEW

The task of generating frequent item sets is the fundamental step for many data mining applications, such as mining association rules, correlations, sequential patterns [3], associative classification [4] etc. Many algorithms have been proposed to find Maximal Frequent item-sets. Some of them are MaxEclat & MaxClique [5], Pincer search [6], Maxminer [7] and Depth project [8].

Trnka uses Data Mining Methods for Market Basket Analysis [9]. This paper describes how Market Basket Analysis can be implemented to Six Sigma methodology. This paper illustrates how market basket analysis increases the performance of sigma.

Mining Interesting Rules by Association and Classification Algorithm is put forth by Yanthy et al. [10] The main intention in data mining is to disclose hidden knowledge from data and several techniques have been suggested so far.

Market Basket Analysis Based on Text Segmentation and Association Rule Mining is proposed by Xie et al. [11]. In this paper, an innovative market basket analysis method is formulated that generates association rules on the basis of items' internal characteristics.

Market Basket Analysis of Library Circulation Data is proposed by Cunningham et al. [12]. In this paper a-priori market basket tool is used to detect subject classification grouping that co-exist in transaction records of books borrowed from a library. This data can be utilized to recommend students other documents relevant to their information need.

Application of Data Mining Techniques for Library Management Information System is given by Bikash Mukhopadhyay and Sripati Mukhopadhyay [13]. This paper illustrates how data mining techniques can be applied to a library management system with the help of artificial intelligence.

Incremental Mining on Association Rules given by Wei-Guang Teng and Ming-Syan Chen explains that record based databases are dynamic [14]. This paper illustrates several algorithms that have been developed for generating precise association rules efficiently and effectively on such dynamic databases.

3. PROBLEM DEFINITION

There are numerous transactions taking place in a library every day which are recorded in a database. In this paper we take some of those numerous transactions, and apply Apriori algorithm over them so as to find out the frequent book items and come up with the association rules on this book items so as to predict the book borrowing behavior of students. It will offer librarians or managers sufficient information to make right decisions. It also describes how incremental mining is incorporated by adding five more transactions to the original set of ten transactions and then compares the association rules generated in two different sets.

4. RESEARCH METHODOLOGY

This section illustrates the general methodology followed during the research. It explains the database design, assumptions etc.

4.1 Database Design

The proposed system is expected to complement the traditional library management software by allowing it to be used for discovering the usage pattern of the books and journals as developed over the years based on the actual transaction data. Thus, in the automated library transaction management system proposed here, there is an online transaction system that stores the current information regarding every transaction of the library. The system stores the data in an operational database. Also these transactions are backed up in a Data Warehouse which stores subject oriented, time variant, non volatile data. These data are then extracted by the data mining tools for knowledge discovery. The features that are incorporated while designing this library transaction management system are:

• Keeping track of the books that are issued to the members.

• Faster mining and retrieval of information.

• Reduced work load of employee.

4.2 Approach Taken For Problem Solving

Database will be created with particular set of domains, with each domain comprising of a set of books. The different books might be represented as <P1, P2 … PN>
The domains considered in this paper are:
• Programming language
• Database Management System (DBMS)
• Data Structures

• Software Engineering

This allows us to identify set of books from same or different domains which are borrowed together by generating association rules from the various transactions. Here association rules are initially generated for ten transactions then an additional set of five transactions are incremented to the original set. Association rules are generated again for the total set of fifteen transactions and the results are compared

4.3 ASSUMPTIONS

1. Books of only four domains are taken which are:
2. One transaction represents the set of books issued by one student at a time.
3. C and C++ are from programming, DB1 and DB2 are database books.
4. The transactions are designed on the basis of common assumptions such as:
a. C and C++ books are generally issued together.
b. .NET and Oracle books are issued together frequently
5. Minimum support = 50%
6. Minimum confidence = 80%

5. RESULTS AND DISCUSSION

This section illustrates how Apriori algorithm is implemented on a Library transactional database to find out first the frequent data item-sets. These item-sets denote the books that are issued frequently by the students. then using the minimum support and confidence association rules are generated which indicate which books are generally borrowed together.
Incremental mining is incorporated by adding five more transactions to the original set of ten transactions and a new set of frequent item-sets and association rules are generated.

This section gives the comparison of the two sets of frequent item-sets and association rules generated.

5.1 Database Design

There are two databases considered in this paper. The first is the initial database containing ten transactions. Every transaction denotes the books issued by a student at one particular instant.
The initial transactional database is represented by the following table.

Table 1 Original transactional database

T ID	Item set
1.	DS,C,.NET,DB1,DB2
2.	.NET,DB1,C,DS,DB2
3.	SW,C++,C,.NET,DB1
4.	SW,C,DS
5.	.NET,SW,DB1,C,DS
6.	C,.NET,DB1,SW,DS,DB2
7.	DB2,DS,C,C++,DB1
8.	DB2,DB1,SW
9.	DB1,C,C++,DS,.NET
10.	.NET,DB1,DB2,SW

The second table that is considered is the incremented table which is added to the original database after association rules are generated for it. Once this table is added to the original database, Apriori is applied to the new set of fifteen transactions to generate another set of frequent item-sets and association rules.

Table 2 Incremented database

T ID	Item-set
1.	C,C++,DB1
2.	DS,C,SW,.NET
3.	.NET,DB1,C
4.	DB2,DB1,SW,DS
5.	DB1,DS,.NET,C,C++

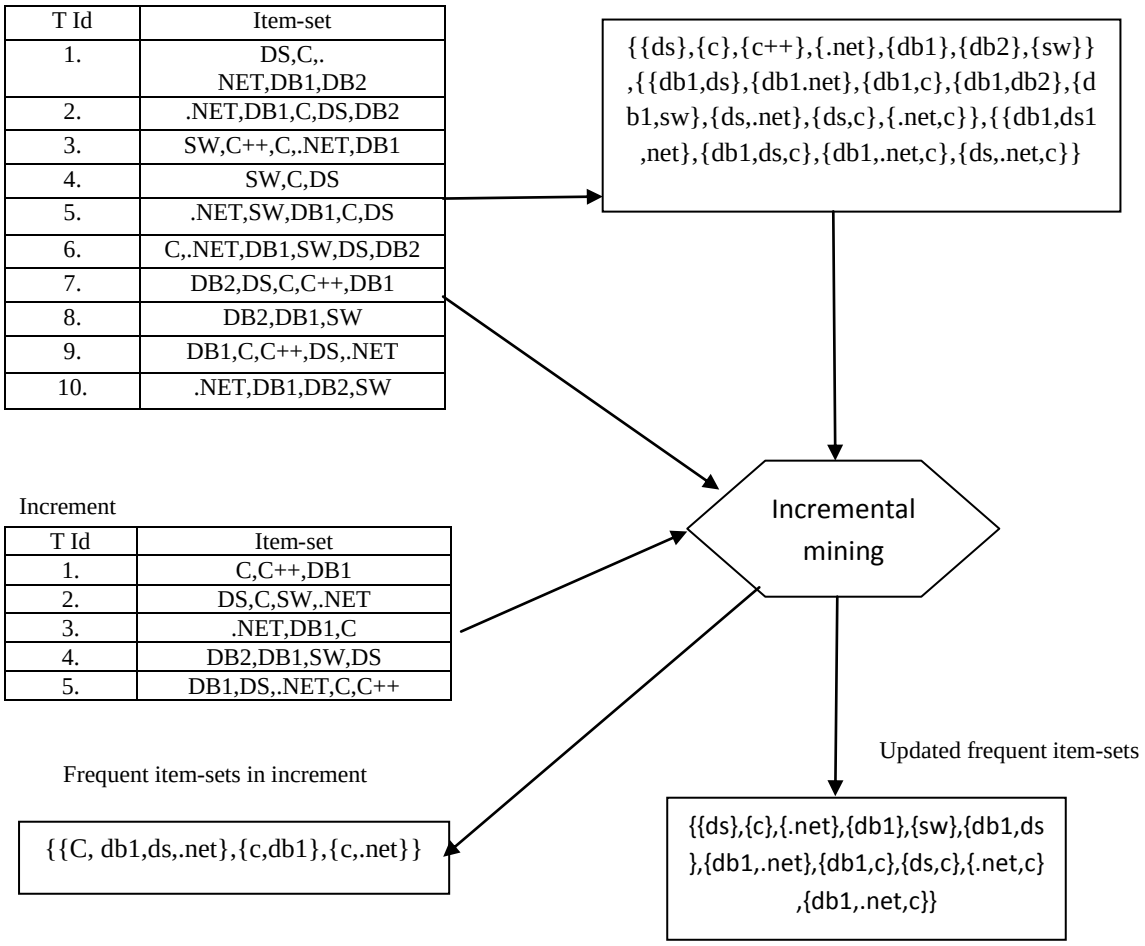

Fig 1: Incremental miming process

5.2Association Rule Generation

The following table illustrates all the association rules generated when Apriori algorithm is applied to the original database of ten transactions using minimum support and confidence of 50 and 80% respectively.

Table 3 Association rules generated

S. No	Rule Generated	Confidence (%)
1.	{DB1,C} ->DS	85.5
2.	{DB1,C} ->.NET	85.5
3.	{DB1,C,DS} ->.NET	83.3
4.	{DB1,C,.NET} ->DS	83.3
5.	{DB1,DS} ->{C,.NET}	83.3
6.	{DB1,DS} ->.NET	83.3
7.	{DB1,DS} ->C	100
8.	{DB1,DS,.NET} ->C	100
9.	{DB1,.NET} ->C	85.5
10.	{C} ->DS	87.5

The rule 1 for example explains that if a student issues DB1 and C book he will also borrow DS book with a confidence or probability of 85.5%. Since the confidence is set to 80% every rule with confidence greater than this is accepted.

5.3 Incremental Mining

Implementation of incremental mining is done, by adding five new transactions to the original transactional database, then association rules are generated for the new database of fifteen transactions. This addition reduces the number of frequent item-sets. The process can be easily explained as shown in figure 1.

11.	{C} ->DB1	87.5
12.	{DS,C} ->DB1	85.7
13.	{C,.NET,DS} ->DB1	100
14.	{C,.NET} ->{DS,DB1}	83.3
15.	{C,.NET} ->DS	83.3
16.	{C,.NET} ->DB1	100
17.	{DS} ->C	100
18.	{DS} ->{DB1,C}	85.7
19.	{DS} ->DB1	85.7
20.	{DS,.NET} ->{DB1,C}	100
21.	{DS,.NET} ->{C}	100
22.	{. NET} ->C	85
23.	{.NET} ->{DB1}	100
24.	{. NET} ->{DB1,C}	85
25.	{DB2 } ->{DB1}	100
26.	{SW}->DB1	83.3

Once incremental mining is applied a new set of association rules are generated using the Apriori algorithm with the

same values of minimum support and confidence. It is seen that the number of rules generated have also decreased. This set of rules is illustrated in the table given below.

5.4 Discussion of Results Obtained

From the above illustrations it is clear that after the implementation of incremental mining the association rules generated were decreased to almost one-third even when the support and confidence are kept same.

Also the rules generated after the database is incremented, is the subset of rules generated in the initial step.

S. No	Rules Generated	Confidence (%)
1.	{DB1,C}-> .NET	80
2.	{DB1,C.DS}-> .NET	85.71
3.	{DB1,DS}-> C	87.5
4.	{DB1,DS,.NET}-> C	100
5.	{DB1,.NET}-> C	88.8
6.	{C}->DB1	83.3
7.	{C,NET}->DB1	88.8
8.	{.NET}->C	90

Table 4 Association rules generated

6. IMPLEMENTATION

Program was written in .NET which implemented Apriori algorithm and generated the corresponding association rules. Subsequently this process was repeated for implementing incremental mining. The new rules generated were compiled and verified.

6.1 Abbreviations of Book Titles

The table below illustrates the abbreviations that are supposed in the hand computation and in the program implementation.

6.2 Snapshots of the Project

The first snapshot represents the first form of the project through which the transactions of the database are fed into the Apriori algorithm for generation of association rules. Also the minimum support and confidence for finding frequent item-sets and generating association rules is entered through this form.

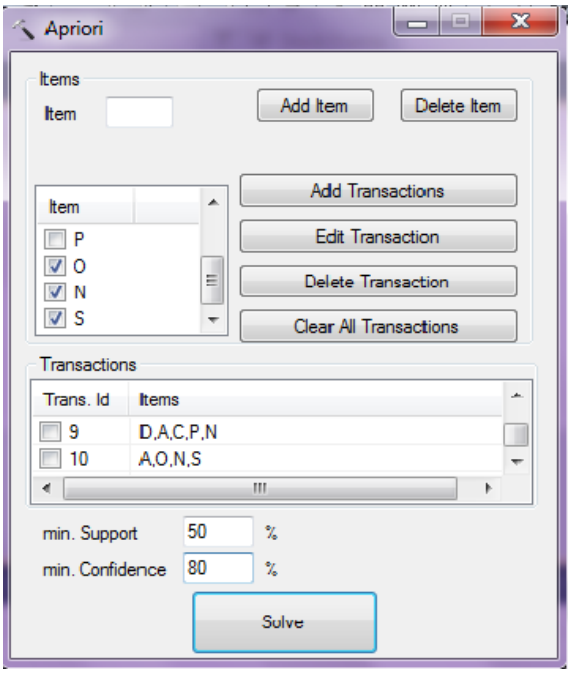

Fig 2: Snapshot 1-Feeding of transactions

The second snapshot represents the output form which generates frequent item-sets, closed item-sets, maximal item-sets and the association rules for the set of ten transactions that are fed into the Apriori algorithm. Here the frequent item-sets and their support is mentioned and also a set of twenty-six rules along with their confidence are shown. The results are absolutely correct as they match the results obtained by hand computation for the same set of transactions.

Table 5 Abbreviations Used

Book Title	Hand computation abbreviations	Program abbreviations
Let Us C	C	C
Programming in c++	C++	P
SQL and PL/SQL	DB1	A
Database Systems	DB2	O
Data structures through C	DS	D
Software Engineering	SW	S
.NET Framework	.NET	N

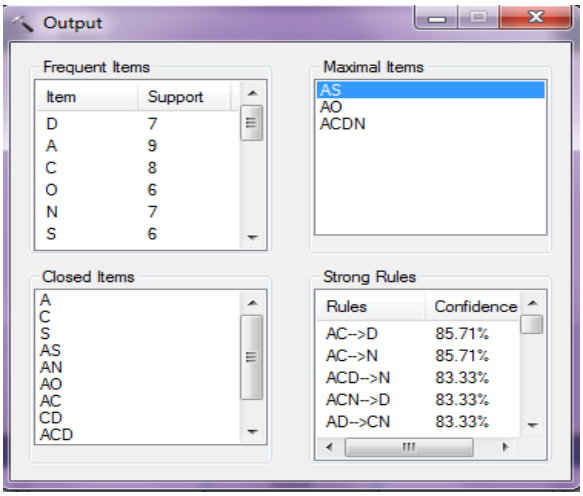

Fig 3: Snapshot 2-Output Form

The third and the last snapshot shows the output form after five more transactions have been added to the original set of ten transactions to implement incremental mining. It can be observed from this form that the number of frequent item-sets, maximal items,closed items and association rules generated have decreased. And the set of association rules in this case is a subset of the first set of rules generated.

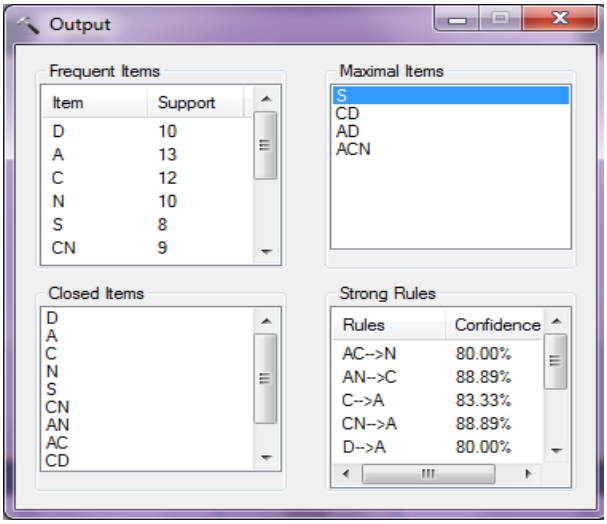

Fig 4: Snapshot 3-Output Form showing incremental mining results

7. CONCLUSION AND FUTURE WORK

The paper first required the study of the University Library to understand the organization and working of the library. Then data mining and association rules were studied and Apriori algorithm was selected to implement decision support system in the library.

Later the database design was selected to constitute only of transaction ID and item-sets and only four domains are selected. On this database then Apriori algorithm was implemented and association rules above minimum support and confidence were selected and stored in tabular form. To the original set of ten transactions five more transactions were added and association rules were generated to see effect incremental mining.

This work is easily extensible. Some of the directions of future work are:

1. Scalability into a Data Mart Application.
2. Use of Cluster Analysis to segregate library users based on a common borrowing and usage profiles.
3. Detection of Outliers in usage patterns of the library.
4. Prompt based summaries for early warning of defaulters.
5. Cluster analysis to determine usage patterns of books and other technical literature in the library.
6. Intelligent inventory management using forecasting and other data mining techniques.
7. Develop a recommendation system that can be used for personal prediction according to a user's borrowing profile.
8. Stacking of books according to borrowing pattern.
9. Selection of books for development of departmental libraries for a university.

8. REFRENCES

[1] Han, Jiewai, Kamber, Micheline and Pei, Jian (2005). Data Mining: Concepts and Techniques. Morgann Kaufmann series in Dara Management Systems

[2] Sarma, PKD and Roy, Rahul. A Data Warehouse for Mining Usage Pattern in Librarty Transaction Data. Assam University Journal of Science & Technology: Physical Sciences and Technology Vol. 6 Number II 125-129, 2010

[3] Agrwal, Rakesh and Srikant, Ramakrishnan. (1994) Fast Algorithms for Mining Algorithms. Proceedings of the VLDB Conference, Santiago, Chile.

[4] Liu et. al. (1998). Integrating Classification and Association Rule Mining. KDD-98, New York, Aug 27-31.

[5] Zaki et. al. (1997). Technical Report: New Algorithms for Fast Discovery of Association Rules. University of Rochester, Computer Science Department.

[6] Lin, D. and Kedem, Z.M. (1998) Pincer-Search: A New Algorithm for Discovering the Maximum Frequent Set. Proceedings of the 6th International Conference on Extending Database Technology (EDBT), Valencia .pp. 105-119.

[7] Bayardo, R.J. (1998) Efficiently mining long patterns from databases in Proceedings of ACM SIGMOD Conference on Management of Data, pp. 85–93, New York, USA.

[8] Agrawal, R., Aggarwal, C., and Prasad, V. (2000) Depth first generation of long patterns. Proceedings of Seventh International Conference on Knowledge Discovery and Data Mining, pp. 108–118.

[9] Trnka, A. (2010). Market Basket Analysis with Data Mining methods.Conference on Networking and Information Technology.

[10] Yanthyet. al. (2009). Mining Interesting Rules by Association and Classification Algorithms. Fourth International Conference on Frontier of Computer Science and Technology. pp. 177-182

[11] Xieet. al. (2010).Market Basket Analysis Based on Text Segmentation and Association Rule Mining.Proceedings of First International Conference on Networking and Distributed Computing.

[12] Cunningham et. al. (1999).Market Basket Analysis of Library Circulation Data. Proceedings of the Sixth International Conference on Neural Information Processing.

[13]Mukopadhyay et.al. Retrieved from http://ir.inflibnet.ac.in/bitstream/handle/1944/226/cali_57.pdf

[14] Tenget. et.al. Retrieved from http://www.cs.ucla.edu/~wwc/course/cs245a/incremental.pdf