Frequent Itemsets Mining for Book Renting System By Using FP-Growth

Sandar Tun,
Computer University (Pyay)
Sandar316@gmail.com

Abstract

Frequent itemsets mining leads to the discovery of associations and correlations among items in large transactional or relational data sets. Frequent itemsets mining is market basket analysis. This paper presents to analyzes books renting habits by finding associations between the different items that borrowers place in their shopping (renting) baskets. In the proposed system, FP-growth (Frequent Pattern growth) is used to find the frequent itemsets without candidate generation. FP-growth is an order of magnitude faster than Apriori for no candidate generation, no candidate test, the use compact data structure, the elimination of repeated database scan and that basic operation is counting and FP-tree building.

Keywords: Data mining, Association Rule, FP-Growth

1. Introduction

Data mining refers to extracting or mining knowledge from large amounts of data. It is also referred to as edge discovery in databases. Many other terms carry a similar or slightly different meaning to data mining such as, knowledge mining from data, knowledge extraction, data/ pattern analysis, data archaeology, and data dredging. Many people treat data mining as in synonym for another popularly used term Knowledge Discovery from Data or (KDD). It aims at extracting previously unknown and potentially useful knowledge from large sets of data. Data mining is the task of discovering interesting patterns from large amounts of data where the data can be stored in databases, data warehouses, or other information repositories. Data patterns can be mined from many different kinds of databases, such as relational databases and transaction databases. Interesting data patterns can also be extracted from other kinds of information repositories, including spatial, text, multimedia, and legacy databases, and the World Wide Web. Data mining systems can be classified according to the kinds of databases mined, the kinds of knowledge mined, the techniques used, or the application adapted. Data mining functionalities include the discovery of concept\class descriptions, association, classification, prediction, clustering, trend analysis, deviation analysis, and similarity analysis. Characterization and discrimination are forms of data summarization [2].

Data mining involves an integration of techniques from multiple disciplines such as database and data warehouse technology, statistics machine learning, high performance computing, pattern recognition, neural networks, data visualization, information retrieval, image and signal processing, and spatial or temporal data analysis. One of the most important data mining problems is discovery of frequently occurring patterns in sequential data. Application areas for this problem include analysis of telecommunication systems, discovering frequent buying patterns, analysis of patients' medical records, etc.

Data Mining is the process of from digging or gathering information from various database .This include data from point of sale transactions, credit card purchases, online forms which are just a few of the many things that some of the large companies dig to find out more about their clients. Several organizations have collected massive amounts of such data. These data sets are usually stored on tertiary storage and are very slowly migrating to database systems. One of the main reasons for the limited success of database systems in this area is that current database systems do not provide necessary functionality for a user interested in taking advantage of this information [5].

Finding frequent pattern plays the important role in mining association, correlations, and many other interesting relationships among data. Recently, frequent item mining is a natural application of data mining techniques for Market Basket Analysis. Frequent pattern mining searches for recurring relationships in given data set. It also leads to the discovery of associations and correlations among items in large transactional or relational data sets. With massive amounts of data continuously being collected and stored, many industries are becoming interested in mining such patterns from their databases. This discovery of

interesting correlations relationships among huge amount of business transaction records can help in many business decision making processes. This paper introduces the problem of “mining" a large collection of basket data type transactions for association rules between sets of items with some minimum specified confidence, and presents a FP-growth algorithm for this purpose.

2. Related Work

The discovery of association relationship among huge amounts of data is useful in selective marketing, decision analysis, and business management. Ascociation rule mining consists of first finding frequent itemsets (sets of itmes, such as A and B, satisfying a minimum support threshold, or percentage of the task-relevant tuples), from which strong association rules in the form of A⇨B are generated. These rules also satisfy a minimum confidence threshold (a precified probability of satisfying B under the condition is A is satisfied). Mining frequent patterns or itemsets is a fundamental and essential problem in many data mining applications [5].

These applications include the discovery of association rules, strong rules, correlations, sequential rules, episodes, multidimensional patterns, and many other important discovery tasks. Algorithms for extracting this basis and for reconstructing all association rules shows the results of experiments carried out on real datasets and it show the usefulness of each approach. Frequent patterns mining plays an essential role in many datamin tasks, such as mining association rule, correlations, causality, sequential patterns, episodes, multi-dimensional patterns, max-patterns, partial periodicity, and emerging patterns. Frequent pattern mining techniques can also be extended to solve many other problems, such as iceberg-cube computation classification. Thus, effective and efficient frequent pattern mining is an important and interesting research problem. For many applications, it is difficult to find strong association among data items at low or primitive levels of abstraction due to the aparsity of data in multidimensional space. Strong association discovered at high concept levels many represent common sense knowledge. However, what may represent common sense to one user may seem novel to other. Therefore, data mining systems should provide capabilities to mine association rules at multi levels of abstraction and traverse easily among different abstraction space.

3. Mining Association Rule

Association Rule means "Customer who bought item A also bought item B". Association Rule Mining also means "To extract such rules from transaction database". Association rule mining finds interesting association or correlation relationships among a large set of data items. Data mining is the process of automatically discovering useful information in large data repositories. Data mining techniques are deployed to scour large database in order to find novel and useful patterns that might otherwise remain unknown. Data mining tasks are generally divided into two major categories:

- Predictive tasks: perform inference on the current data in order to make predictions.
- Descriptive tasks: characterize the general properties of the data in the database.

Association analysis is used to discover patterns that describe strongly association features in the data. These discovered patterns are typically represented in the form of implication rules or feature subsets. Because of the exponential size its search space, the goal of association analysis is to extract the most interesting pattern in an efficient manner. With massive amounts of data continuously being collected and stored, many industries are becoming interested in mining association rules from their databases. The discovery of interesting association relationships among huge amounts of business transaction records can help in many business decision making processes, such as catalog design, cross marketing, can loss-leader analysis.

A typical example of association is market basket analysis. This process analyzes customer buying habits by finding associations between the different items that customers place in their “ shopping baskets”. The discovery of such association can help retails develop marketing strategies by gaining insight into which items are frequently purchased together by customers. For instance, if customers are buying milk, how likely are they to also buy bread (and what kind of bread) on the same trip to the supermarket. Such information can lead to increased sales by helping retailers do selective marketing and plan their shelf space. For example, placing milk and bread within close proximity may further encourage the sale of these items together within single visits to the store.

4. FP-Growth Mining

Jiawei Han and Micheline Kamber1designed a method that mines the complete set of frequent itemset without candidate generation. This interesting method is called the frequent pattern growth, FP-growth, which adopts a divide-and-conquer strategy as follows: compress the database representing frequent items into a frequent pattern

tree, FP-tree, but retain the itemsets association information, and then dive such a compressed database into a set of conditional database, each associated with one frequent item, and mine each such database separately. FP-Growth finds all the frequent itemsets ending with a particular suffix by employing a divide and conquer strategy to split the problem into smaller sub- problems. For example, suppose we are interested in finding all frequent itemsets ending in e. To do this, we must first check whether the itemset {e} itself is frequent. If it is frequent, we consider the subproblem of finding frequent itemsets ending in de, ce, be and ae. In turn, each of these subproblems are further decomposed into smaller subproblems. By merging the solutions obtained from the subproblems, all frequent itemsets ending in e is found. Mining frequent itemsets without candidate generation consists of two steps [2].

4.2 Step I-Building Frequent Pattern Tree

In the Step I, there are seven sub-steps.

(1) First, create the root of the tree, labeled with "null".
(2) Scan the database D a second time. (First time we scanned it to create 1-itemset and then L).
(3) The items in each transaction are processed in L order (i.e. sorted order).
(4) A branch is created for each transaction with items having their support count separated by colon.
(5) Whenever the same node is encountered in another transaction, we just increment the support count of the common node or Prefix.
(6) To facilitate tree traversal, an item header table is built so that each item points to its occurrences in the tree via a chain of node-links.
(7) Now, the problem of mining frequent patterns in database is transformed to that of mining the FP-Tree.

4.2 Step II- Finding Frequent Itemsets on FP-growth Technique

(1) Start from each frequent length-1 pattern (as an initial suffix pattern).
(2) Construct its conditional pattern base which consists of the set of prefix paths in the FP-Tree co-occurring with suffix pattern.
(3) Then, Construct its conditional FP-Tree & perform mining on such a tree.
(4) The pattern growth is achieved by concatenation of the suffix pattern with the frequent patterns generated from a conditional FP-Tree.
(5) The union of all frequent patterns (generated by step 4) gives the required frequent itemset.

4.3 Case Study

In the following figure, there are four transactions.

Table 1. Book Transaction

Transaction ID	Book Items
T1	AI,DS,SE,MIS
T2	AI,MIS
T3	AI,CG
T4	DS,SE,PL

Book transactions are described with their frequency number at items header table. The minimum support count for this case study is supposed as 2. So, CG and PL are neglected as they have 1 support count. Then the FP-tree is drawn, based on items header table.

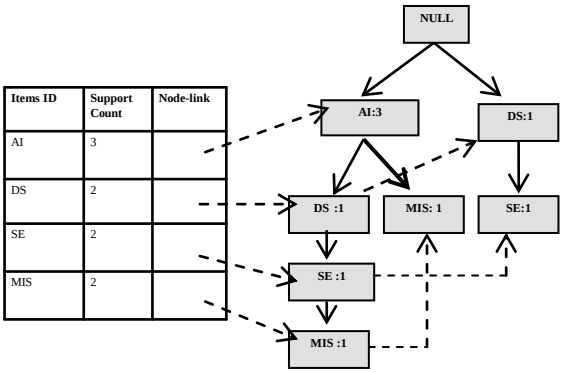


Figure 1. FP-tree from Items Header Table

Table2. Mining FP-tree by Creating Conditional Sub Pattern Bases

Item	Conditional Patterns Base	Conditional FP-Tree	Frequent Pattern Generated
MIS	{(AI,DS,SE:1), (AI:1)}	{AI:2}	{AI,MIS:2}
SE	{(AI,DS:1), (DS:1)}	{DS:2}	{SE,DS:2}
DS	{(AI:1)}	-	-
AI	-	-	

Output association rules from Frequent-Pattern Generated process are:

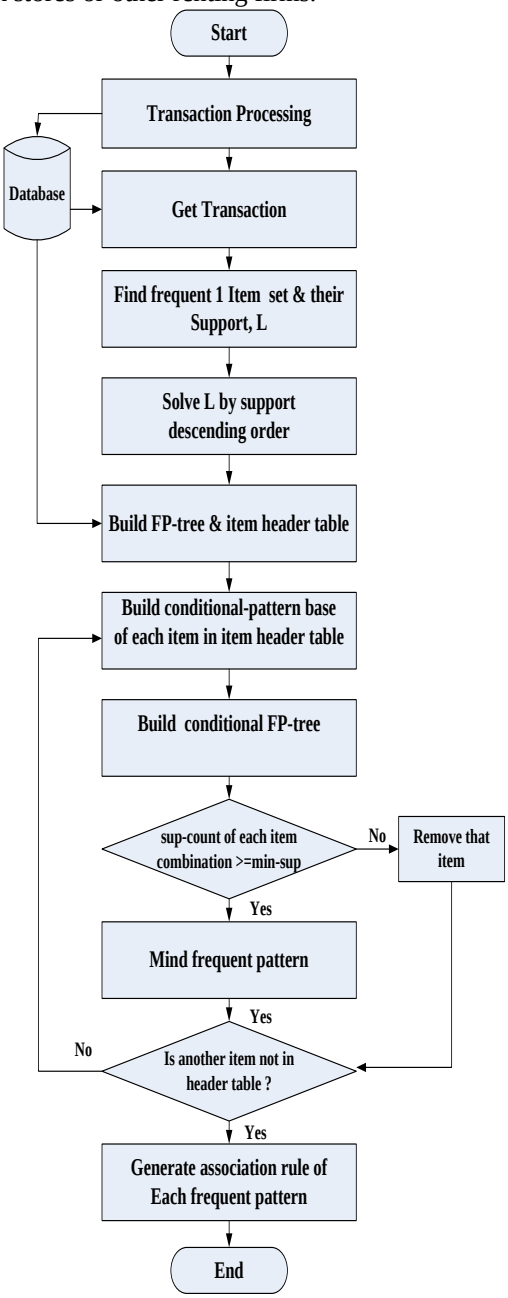
AI→MIS
SE→DE

5. Architecture of the System

This paper describes the design and implementation of frequent itemsets mining of book renting system. The main proposes of this system are to help renters in analyzing the data to learn the renting behavior of their customers and to gain insight into which items (books) are frequently

rented together by borrowers. To draw FP-tree, the system accepts a minimum support count from the user. The itemsets to draw FP-tree can change if the minimum support count changes.

In this system, the book types are the items and there are mainly 15 book types. The final results of this system are association rules generated from frequent patterns. This system also is similar to the library system. The owners can see whether the books are returned and the list of book renting at Show Rent List menu. The system displays all books and the user can choose what book he wants to borrow. In this system, the borrowers can member the book store too. But this system designs to be a prototype for not only the libraries but also book stores or other renting firms.



.Figure 2. Architecture of the System

6. Result of the System

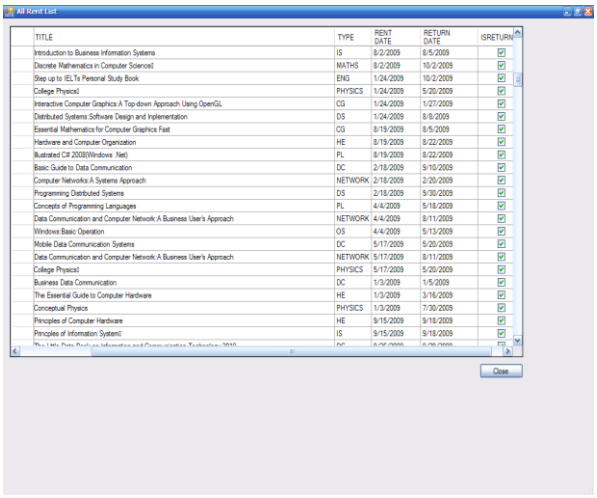


Figure 3. Book Rent List

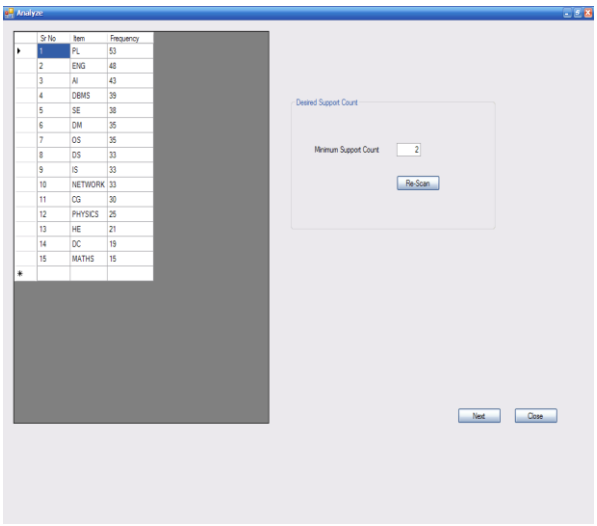


Figure 4. User gives a minimum support count

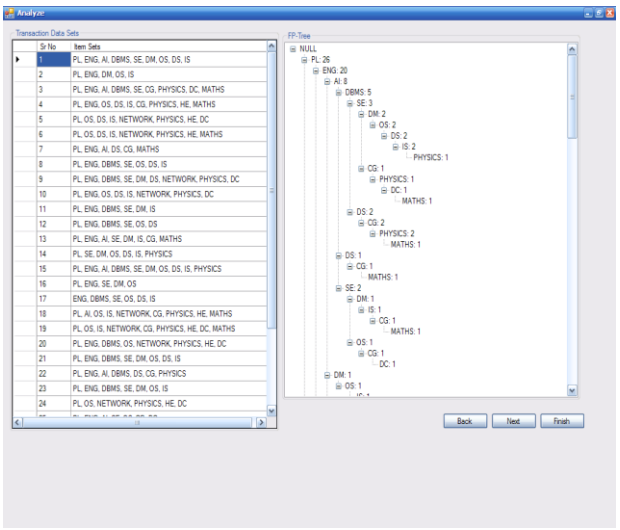


Figure 5. Transaction Data Set and FP-tree

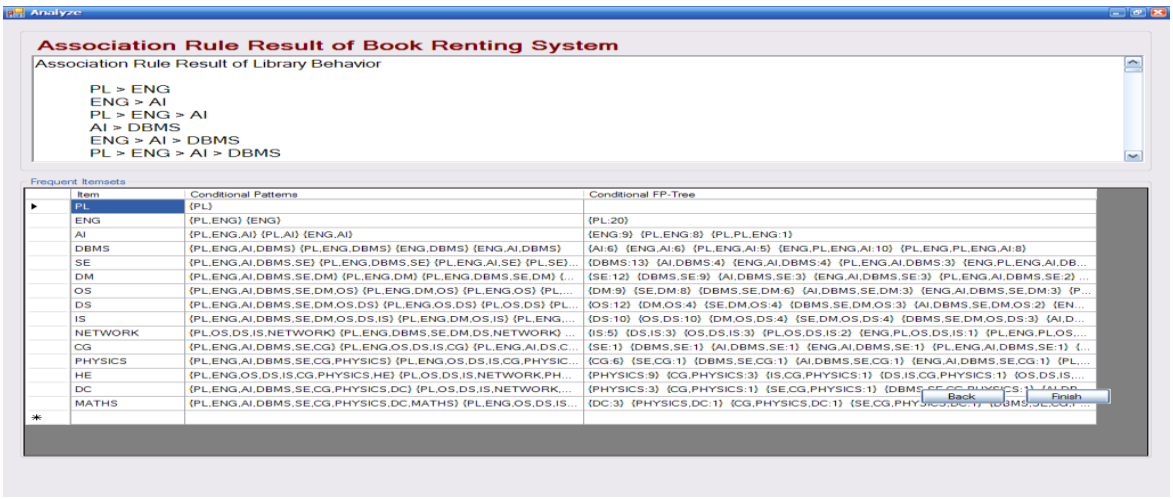


Figure 6. Association Rules of Book Items

7. Conclusion

This system implements the book renting process for the renter and borrowers. The system can guest which books get the highest level of borrowers and they can buy. So, this can extend book renting marketing. The customer can also know which books they like the best match with another books and they rent these books on the matching frequent. This system use FP-growth to generate the output association rule. FP-Growth is an interesting algorithm because it illustrates how a compact representation of the transaction data set helps to effectively generate frequent itemsets. So, this system can be used in decision making process in a reasonable time frame to solve for book renting process in a highly competitive world.

8. References

[1]. B.Goethals and M.J.Zaki. "Advances in frequent itemset mining implementations": introduction to fimi03. *In proceeding of the 1st IEEE ICDM* Workshop on Frequent Itemset

[2]. J.Han, M. Kamber, "*Data Mining Concepts and Techniques"*, Simon Fraser University, 2001 byAcadamic Press.

[3]. "Generating a Condensed Representation for Association Rules", *Journal of Intelligent Information Systems*, 24:1, 29–60, 2005. 2005 Springer Science Business Media, Inc. Manufactured in the Netherlands.

[4]. H.Liu, "*Market Basket Analysis and Itemsets APRIORI Efficient Association Rules"*, Multilevel Association Rule, Post- processing, CSE 572, CBS 598.

[5]. Agrawal, T. Imielinski, and A.Sawmi, "*Mining association rules between sets of items in large databases*." In proc. of the ACM SIGMOD Conference on Management of Data, pages 207-216, May 1993.

[6]. R.Agrawal and R. Srikant, "*Fast Algorithms for Association Rule Mining"*, IBM Almaden Research Center. 1994.

[7]. Rakesh Agrawal , Tomasz Imielinski , Arun Swami , " *Mining Association Rules between Sets of Items in Large Databases*". IBM Almadern Research Center, 650 Harry Road , San Jose CA 95120