Analysis of Book Sales Transaction using Association Rules Mining

Ei Phyu Zin, Nilar Aye
University of Computer Studies, Yangon
eiphyuzin2008 @gmail.com, nilaraye9 @gmail.com

Abstract

Data Mining is the process of analysis of raw data in the database and synthesizing it into information that is useful for effective decision making. Association rule mining finds interesting association or correlation relationships among a large set of data items. The discovery of interesting association relationships among huge amount of business transaction records can help in many business decision making processes.

A typical example of association rule mining 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 retailers develop marketing strategies by gaining insight into which items are frequently purchased together by customers.

This thesis, is intended to develop a system for market basket analysis on Book Store which will generat strong efficient rules among itemsets with use of Apriori Algorithm.This system will be implemented the Java Programming language.

Keyword: Data Mining, Association Rule Mining, Market Basket, Apriori algorithm.

1. Introduction

Data mining is the process of discovering interesting knowledge from large amount of data stored either in database, 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 database 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.

Data mining is the process of from digging and gathering information from various databases. This includes 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. The information is used to find out how major of the clients behavior, or what makes them irritated or simply how can they make the life of the client happier.

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.

Frequent item-set mining 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. The discovery of interesting association relationships among huge amount of business transaction records can help in many business decision making processes such as catalog design, cross-marketing, and customer shopping behavior analysis.

2. Related Work

Association rule mining finds interesting association or correlation relationships among a large set of data items. The discovery of interesting association relationships among huge amount of business transaction records can help in many business decision making process such as catalog design, cross-marketing, and loss leader analysis.

Traditionally, association rules are used to discover business trends by analyzing customer transactions. However, they can also be used effectively to predict web page accesses for personalization. For example, assume that after mining the web access log we discovered an association rule “A and B implies C”, with 80% confidence, where A, B and C are web page accesses. If a user has visited pages A and B, there is an 80% chance that he/she will visit page C in the same session. Page C may or may not have a direct link from A or B. This information can be used to create a link dynamically to page C, from page A or B, so that the user can “click-through” to page C directly. This kind of information is particularly valuable for a web server supporting an e-commerce site to link

the different product pages dynamically, based on the customer interaction.[1]

We are given a large database of customer transactions. Each transaction consists of items purchased by a customer in a visit. Consider a supermarket with a large collection of items. Typical business decisions that the management of the supermarket has to make include what to put on sale, how to design coupons, how to place merchandise of shelves in order to maximize the profit, etc. Analysis of past transaction data is commonly used approach in order to improve the quality of such decisions. Until recently, however, only global data about the cumulative sales during some time period (a day, a week, a month, etc.). was available on the computer, Progress in bar-code technology has made it possible to store the so called basket data that stores items purchased on a per transaction basis. Basket data type transactions do not necessarily consist of items bought together at the same point of time. It may consist of items bought by a customer over a period of time, Examples include monthly purchases by members of a book club or a music club [4].

3. Theoretical background

3.1 Market Basket Analysis

Market basket analysis may be preformed on the retail data of customer transactions at store. The result of analysis can use to plan marketing or advertising strategies, or in the design of a new catalog. In this strategy, items that are frequently purchased together can be placed in proximity in order to further encourage the sale of such items together. If customers who purchase Computer book also tend to buy Technology book at the same time, then placing the technology book display close to tend to the computer book display may help increase the sales of both items.

Market basket analysis can also help retailers plan which items put on sale at reduced prices. If customers tend to purchase computer book and technology book together, then having a sale on computer book may encourage the sale of technology books as well as computer books.

3.2 Association Rule

If the set of items available at the bookstore, then each item has a Boolean variable representing the presence or absence of that item. Each basket can then be represented by a Boolean vector of values assigned to these variables. The Boolean vectors can be analyzed for buying patterns that reflect items that are frequently associated or purchased together. These patterns can be represented in the form of association rules. For example, the information that customer who purchase computer book also tend to buy technology book at the same time is represented in Association Rule below:

Computer book => Technology book [support = 2%, confidence = 60%]

Rule support and confidence are two measure of rule interestingness. They respectively reflect the usefulness and certainty of discovered rules. A support of 2% for Association Rule means that 2% of all the under analysis show that computer book and technology book are purchased together. A confidence of 60% means that 60% of the customers who purchased a computer book also bought the technology book. Typically, association rules are considered interesting if they satisfy both a minimum support threshold and a minimum confidence threshold. Such thresholds can be set by users or domain experts.

3.3 Frequent Itemsets and Association Rules

Let $I = \{I_1, I_2, \ldots\ldots, I_m\}$ be a set of items. Let D, the task-relevant data, be a set of database transactions where each transaction T is a set of items such that T€I. Each transaction is associated with an identifier, called TID. Let A be a set of items. A transaction T is said to contain A if and only if $A \subseteq T$. An association rule is an implication of the form A => B, where $A \subset I$, $B \subset I$ and $A \cap B = \phi$. The rule A => B holds in the transaction set D with support s, where s is the percentage of transactions in D that contain $A \cup B$.

This is taken to be probability, P ($A \cup B$). The rule A => B has confidence c in the transaction set D if the c is the percentage of transactions in D containing A that also contain B. This is taken to be the conditional probability, P (B\A). That is,

Support (A => B) = P ($A \cup B$)

Confidence (A => B) = P (B\A)

In general, the association rule mining can be viewed as a two-step process:

1. Find all frequent itemsets: By definition, each of these itemsets will occur at least as frequently as predetermined minimum support count, min_sup.
2. Generate strong association rules from the frequent itemsets: By definition, these rules must satisfy minimum support and minimum confidence.

3.4 Association Rule Mining

Association Rule mining is a data mining task that discovers relationships among items in a transactional database. Association rule analysis is the task of discovering association rule that occur frequently in a given dataset. Association rule mining consists of first finding frequent item (set of

items, 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. [2]

Association rule mining or induction is commonly used in market basket analysis to find items frequently bought together by shoppers. The first algorithm for mining frequent item sets is the Apriori was used for market basket analysis.

For example, if amazon.com discovers that shoppers who buy the book " Data Mining: Concept and Techniques" usually buy another book " Data Mining Practical Machine Learning Tools and Techniques with java implementations", they can arrange an offer for an offer package of these 2 books to increase their competitiveness in safe. The rules are induced from items that are most frequently occurred together, known as the frequently item set A rule "Buy (A) ^ Buy(B).Buy (C)" indicates that a customer who buys item A and item B buys item C, with the interestingness of this rule measured from probabilities of support and confidence.

3.5 Association Rule Mining: A Road Map

Association rules can be classified in various ways, based on the following criteria:

1. Based on the types of values handled in the rule: Boolean association rule and Quantitative association rule
2. Based on the dimension of data involved in the rules: Single dimensional association rule and Multi-dimensional association rule
3. Based on the levels of abstractions involved in the rule: Single level association rule and Multi-level association
4. Based on various extension to association mining: Max patterns and frequent closed itemset.

4. Methodology

4.1 Association Rule Mining Algorithm

This section will explain method for mining the simplest form of association rules, single-dimensional, single-level, Boolean association rules, such as those discussed for market basket analysis. Apriori is an influential algorithm for mining frequent itemsets for Boolean association rule. First, Apriori algorithm, a basic algorithm finding frequent itmesets will be explained: Generating Association Rules from frequent itemsets. Once the frequent itemsets from transactions in a database D have been found, it is straight forward to generate strong association rules from them (where strong association rules satisfy both minimum support and minimum confidence).

4.1.1 Apriori Algorithm

Apriori algorithm is an efficient association rule mining algorithm that explores the level-wise mining. Apriori property: all non empty subsets of a frequent itemset must also be frequent. The Apriori algorithm is a popular correlation based data mining kernel [5].

Table 1. Notation

k-item set	An itemset having k items
L_k	Set of large k-itemsets, whose supports is larger than user specified minimum support
C_k	Set of candidate k-itemsets, which is potentially large itemset.

However, it is a computationally expensive algorithm and the running times can stretch up to days for large database, as database sizes can reach from Gigabytes and computation require multiple passes. Figure 1 gives an overview of the algorithm, using the notation given in Table 1.

Apriori employs an iterative approach know as a level-wise search, where k-itemsets are use to expore (k+1)-itemsets. First, the set of frequent 1-itemsets is found by scanning the database to accumulate the count for each item, and collecting those items that satisfy minimum support. The resulting set is denoted L_1. Next, L_1 is used to find L_2, the set of frequent 2-itemsets, which is used to find L_3, and so on, until no more frequent k-itemsets can be found. The finding of each L_k requires one full scan of the database.

```
Input D, a database of transactions;
min_sup, the minimum support count threshold.
Output: L, frequent itemsets in D.
Method:
L1 = find frequent 1-itemsets(D);
For (k==2; Lk-1 ≠ 0, k++{
        Ck = apriori_gen(Lk-1);
        For each transaction t Є D
                Ct = subset (Ck,t)
                For each candidate c Є C1
                        c.count++;
        }
        Lk = (cЄ Ck)\count ≥ min_sup)
}
Return L = U k L k
```

Figure 1. Apriori Algorithm

In the first pass (pass I), support-count for each item is counted by scanning the transaction

database. Here after we prepare a field named support-count for each itemset, which is used to measure how many times the itemset appeared in transactions. Since itemset here contains just single item, each item has a support-count field. All the items which satisfy the minimum support are picked out. These items are called large 1-itemset (L_1). The second pass (pass 2), the 2-itemset are generated using the large 1-itemsets which is called the candidate 2-itemsets (Cs). Then the support-count of the candidate 2-itemsets is counted by scanning the transaction database. At the end of scanning the transaction data, the large 2-itemsets (L2) which satisfy minimum support are determined. [3]

The following denotes the k-th iteration, pass k.

(1) **Generate candidate itemset**: The candadite k-itemsets (C_k) are generated using large (k-1) itemsets (L_{k-1}) which were determined in the previous pass.
(2) **Count Support**: The support-count for the candidate k-itemsets are counted by scanning the transaction database.
(3) **Determine large itemset**: The candidate k-itemsets are checked for whether they satisfy the minimum support or not, the large k-itemsets (Lk) which satisfy the minimum support are determined.
(4) The procedure terminates when the large itemset becomes empty. Otherwise k:= k+1 and goto "1".

4.2 Apriori Candidate Generation

The procedure for generating candidate k-itemsets using (k-1) itemsets is as follows: Given a large (k-1) itemset, we want to generate a superset of the set of all large k-itemsets. Candidate generation occurs in two steps. First, in the join step, join large (k-1) itemset with (k-10 itemsets. Next, in the prune setp, delete all of the itemsets in the candidate k-itemset where some of the (k-1) subset of candidate itemsets are not in the large (k-1) itemsets.

4.3 Rule Generation

This section describes how to extract association rules efficiently from a given frequent itemsets. Each frequent k-itemsets, Y, can produce up to 2k-2 association rules, ignoring rules that have empty antecedents or consequence (0 → Y or Y →0). An association rules can be extracted by partitioning X, such that X →Y. X satisfies the confidence threshold. Note that all such rules must have already met the support threshold because they are generated from a frequent itemsets.

4.4 Rule Generation in Apriori Algorithm

The Apriori algorithm uses a level-wise approach for generating association rules, where each level corresponds to the number of items that belong to the rule consequent. Frequent itemsets do no mean association rule. One more step is required to convert these frequent itemsets into rules. Association Rules can be found from every frequent itemset X as follows:

For every non-empty subset A of X

1. Let A = X A

2. A ==>B is an association rule if

Confidence (A=>B)>= min confidence where Confidence (A=>B)= support (AB)/support (A) and Support (A=> B) = support (AB)

All the high supprt rules that have only one item in the rules consequent are extracted.

5. Proposed System Architecture

This system will implement the analysis of Book store transaction using association rules mining. The book store retailer wants to know the information of book sales by customer, time period, and store layout etc. This information uses the results of plan marketing and advertising, design of retail layout and design of retail layout. The book items that are frequently purchased together can be placed in proximity in order to further encourage the sale of such items together. This system produces the analysis report for different books on time dimension.

This system develops a method for analyzing buying pattern of the customers in the book store. Association Rules will be used in the data mining process. Association rule mining searches for interesting relationships among items in a given data set. Suppose in a book store, a owner would like to learn more about the buying patterns of the customers. A market basket analysis can be preformed on the retail data of customer transaction of the book store.

Let D be a database of transaction. Each transaction consists of a transaction identifier and a set of items $\{i_1, i_2, \ldots, i_n\}$ selected from the universe I of all possible descriptive items. Table 1 shows the transaction data of customer buying books. There are 7 transactions in this database, that is |D| =7.

Table 2. Transaction Data

Transaction ID	List of items

T001	Zaykwat, AhLinTan, FirstEleven
T002	Pututu,Shwethwe,Tayza,Olympic
T003	Internet,Olympic,Zaykwat, AhLinTan
T004	Modern,BiEleven,First Eleven,
T005	7Days News, AllStar, Pyi Myanmar
T006	Internet, Modern, Pututu, Olympic
T007	Zaykwat, 7DaysSport, AhLinTan

To illustrate the Apriori algorithm for finding frequently item sets in D, the following figure is used.

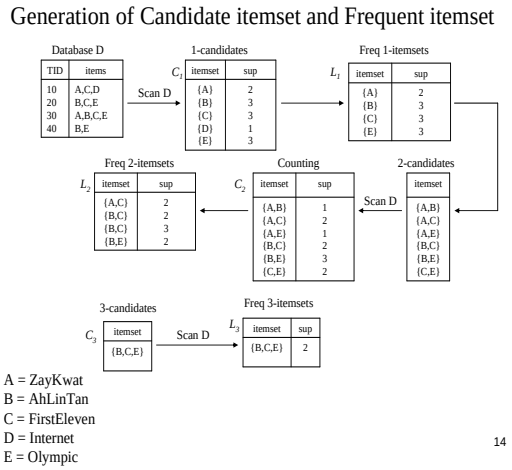

Figure 2. Generation of Candidate itemset and frequent itemsets.

In Figure 3, it shows the process flow of the system. These are step by step processing to generate association rules. Firstly, support count for each item is found. Then it is compared with min-sup count. Items less than min-sup count are removed and others are go on processing. After that, sup-count for item pairs is got. We then again compare each of them with min-sup count and remove pairs which are less than min-sup count. After finishing these processing, we can go on to generate association rules. Finally association rules come out as output analytical result.

5.1 Experimental Result

We examined with the rule mining algorithm using the book store data obtained from different book store. Figure 4 shows the book list of the system.

101	7 Days
102	Ah Lin Tan
103	Zaykwat
104	Family Health
105	Hight light

Figure 4. Example data of book list

Rule 1	101,104 (90%)
Rule 2	102,105 (88%)
Rule 3	101,103 (80%)
Rule 4	104, 105 (76%)
Rule 5	103,105 (75%)

Figure 5. Example Output for Book Association Rules

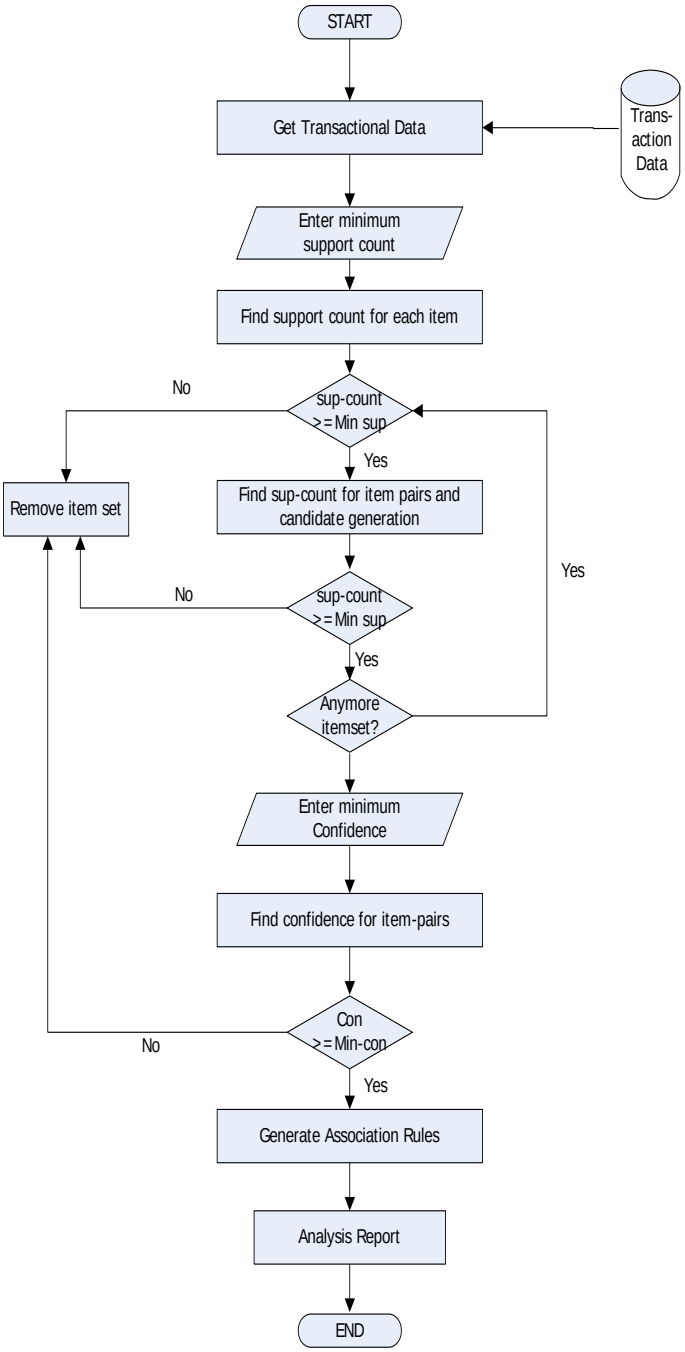

Figure 3. Process Flow of the system

6. Conclusion

Association rule mining is to discover buying patterns such as two or more items that are bought together often from the database. It is popular and useful in a wide variety of situations such as electronic commerce, web mining and bioinformatics. The system will be implemented for finding frequent itemsets and strong association rules which satisfy minimum support and minimum confidence.

References

[1] Borgelt, C and Kruse, R., "Induction of association rules: Apriori implementation. 14th Conference on Computational Statistics, 2002.

[2] Han J., and Kamber M., " Data Mining Concepts and techniques", Academic Press, USA, 2001.

[3] Rakesh Agrawel, Tomasz kmielinski, Arun Swami, " Mining Associaton Rules Between sets of items in large databases", IBM Almadern Research Center, 650 Harry Road, San Jose CA 95120.

[4] Takahiko SHINTANI, Masaru KITSYREGAWA, "Hash based parallel / Algorithms for mining association rules," University of Tokyo, Institute of Industrail Science 3rd Dept, 7-22-1, Roppongi, Minato Tokyo 106 Japan

[5] Pei J., Han J., and Mao R. "Closet:An efficient algorithm for mining frequent closed itemset" DMKD00, USA 2000.