Applied Closet Algorithm for Public Library Book Borrowing Analysis

Seint Wint Thu[1], Hlaing Phyu Phyu Mon[2], Pa Pa Win[3], War War Myint[4]
[1] *University of Computer Studies (Meiktila), Faculty of Information Science*
[2,3,4] *University of Computer Studies (Meiktila), Faculty of Information Science*

***Abstract*-Data mining is an emerging area to discover knowledge from a tremendously large database or data warehouse. Association rule mining is one of the important roles in data mining which produces the rules for the associated itemsets in transactional database and examines user behavior. An interesting method to frequent closed itemset mining without generating candidate itemset called CLOSET is used. The CLOSET algorithm was designed to extract frequent closed itemsets from large databases. CLOSET is an FP-tree-based database projection method for efficient mining of frequent closed itemsets. The system explores the partition-based projection mechanism for scalable mining. It develops a single prefix path compression technique to identify frequent closed sets for books quickly. In this system, the database of book transactions borrowed by borrowers is considered as the applied data and the system is focused on generating the frequent closed book-sets and association rules. The system presents that borrowers are interested which books and authors. The system also presents the interestingness and correlation of books and authors. The system generates strong association rules when the minimum support count and confidence are high.**

***Index Terms*- Data mining, CLOSET, FP-tree-based database projection, association rules**

I. INTRODUCTION

Data mining has attracted a great deal of attention in the information industry and in society as a whole in recent years, due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge. The information and knowledge gained can be used for applications ranging from market analysis, fraud detection, and customer retention, to production control and science exploration.

The rapid growth and integration of databases provides scientists, engineers, and business people with a vast new resource that can be analyzed to make scientific discoveries, optimize industrial systems, and uncover financially valuable patterns. This takes these large data analysis projects, researchers and practitioners have adopted established algorithms from statistics, machine learning, neural networks, and databases and have also developed new methods targeted at large data mining problems [10].

Data mining is one component of the exciting area of machine learning and adaptable computation. The goal of building computer systems that can adapt to their environments and learn from their experience has attracted researchers from many fields, including computer science, engineering, mathematics, physics,neuroscience, and cognitive science.

Out of this research has come a wide variety of learning techniques that have the potential to transform many scientific and industrial fields. Several research communities have converged on a common set of issues surrounding supervised, unsupervised, and reinforcement learning problems [7]. Data Mining is the process of discovering new correlations, patterns, and trends by digging into large amounts of data stored in warehouses. It is related to the subareas of artificial intelligence called knowledge discovery and machine learning. Data mining can also be defined as the process of extracting knowledge hidden from large volumes of raw data i.e. the nontrivial extraction of implicit, previously unknown, and potentially useful information from data [4].

II. THEORETICAL BACKGROUND

The frequent itemsets is the important part of data mining. so some concepts are required to find out frequent itemset. The first number is called the support for the rule. The support is simply the number of transactions that include all items in the antecedent and consequent parts of the rule. (The support is sometimes expressed as a percentage of the total number of records in the database.) The other number is known as the confidence of the rule.

Confidence is the ratio of the number of transactions that include all items in the consequent as well as the antecedent (namely, the support) to the number of transactions that include all items in the antecedent. Lift is one more parameter of interest in the association analysis. Lift is nothing but the ratio of Confidence to Expected Confidence. It is a frequent itemset that is both closed and its support is greater than or equal to minsup. An itemset is closed in a data set if there is no superset that has the same support count as this original itemset. First identify all frequent itemsets. Then from this group find those that are closed by checking to see if there exists a superset that has the same support as the frequent itemset, if there is, the itemset is disqualified, but if none can be found, the itemset is closed. An alternative method is to first identify the closed itemsets and then use the minsup to determine which ones are frequent.

Frequent Itemset Mining can be classified into two methods: Maximal frequent itemset mining and closed frequent itemset mining [3].

III. ASSOCIATION RULE MINING CLASSES

Association rule mining approach can be divided into two classes:

3.1 Bottom Up Approach
3.2 Top Down Approach

3.1 Bottom Up Approach

Bottom Up Approach looks for frequent itemsets from the given dataset that satisfy the predefined constraint. Bottom up approach gets large frequent itemsets through the combination and pruning of small frequent itemsets. The rinciple of the algorithm is: firstly calculates the support of all itemsets in candidate itemsetCk obtained by Lk-1, if the support of the itemset is greater than or equal to the minimum support, the candidate k-itemset is frequent k-itemset, that is Lk, then combines all frequent k-itemsets to a new candidate itemset Ck+1, level by level, until finds large frequent itemsets.

A major challenge in mining frequent itemsets from a large data set is the fact that such mining often generates a huge number of itemsets satisfying the minimum support threshold, especially when min sup is set low. This is because if an itemsetis frequent, each of its subsets is frequent as well. It is useful to identify a small representative set of itemsets from which all other frequent itemsets can be derived such approach is known as Top-Down approach [8, 12].

3.2 Top Down Approach

Top-down Approach looks for more specific frequent itemsets rather than finding more general frequent itemsets. The number of frequent itemsets produced from a transaction data set can be very large. It is useful to identify a small representative set of itemsets from which all other frequent itemsets can be derived. Two such representations are presented in this section are

1) Maximal Frequent Itemset
2) Closed Frequent Itemset

An itemset X is a maximal frequent itemset (or max-itemset) in set S if X is frequent, and there exists no super-itemset Y such that $X \subseteq Y$ and Y is frequent. An itemset X is closed in a data set S if there exists no proper super-itemset Y such that Y has the same support count as X in S. An itemset X is a closed frequent itemset in set S if X is both closed and frequent. The precise definition of closed itemset, however, is based on Relations (1) and (2).Given the functions:

$f(T) = \{i \in I \mid \forall t \in T, i \in t\}$

Which returns all the itemset included in the set of transactions T, and $g(I) = \{t \in T \mid \forall i \in I, i \in t\}$ which returns the set of transactions supporting a given itemset I (its tid-list), the composite function fog is called Galois operator or closure operator. Generator: An itemset p is a generator of a closed itemset y if p is one of the itemsets (there may be more than one) that determines y using Galois

Closure operator: $h(p) = y$

It is intuitive that the closure operator defines a set of equivalence classes over the lattice of frequent itemsets: two itemsets belongs to the same equivalence class if and only if they have the same closure, i.e. their support is the same and is given by the same set of transactions. From the above definitions, the relationshipbetween equivalence classes and closed itemsets is clear: the maximal itemsets of all equivalence classes are closed itemsets. Mining all these maximal elements means mining all closed itemsets [11].

IV. CLOSET ALGORITHM

CLOSET algorithm is a pattern growth method, based on dataset organization. CLOSET algorithm makes the use of the principles of the FP-Tree data structure to avoid the candidate generation step during the process of mining frequent closed itemsets. This work introduces a technique, referred to as single prefix path compression that quickly assists the mining process. CLOSET also applies partition-based projection mechanisms for better scalability. The CLOSET algorithm is a very efficient but highly complex technique. The CLOSET algorithm was designed to extract frequent closed itemsets from large databases. It reduces both the computational and cognitive cost in association rule analysis, by limiting the results to just frequent closed itemsets. These algorithms start with scanning for frequent items. The algorithm divides the frequent items by finding just the frequent closed itemsets [6]. The CLOSET technique continues by recursively mining the subsets of the frequent item closed sets. The algorithm then effectively creates conditional databases of the frequent closed-items separately from the initial transactional database. The actual mechanics of this process can become little complex. It begins by calculating the amount of support for items, and including any item above a minimum support level in a list, defined by the particular data being studied. The list of items meeting these criteria becomes the "f list" of frequently occurring items.

The process of dividing the search space then takes each item and produces a new set for it, excluding each of the previous items. After that, the algorithm populates these sets by searching for items which fulfill the criteria for exclusion. This can be conceived of as creating a number of conditional databases of frequent closed itemsets.

The CLOSET approach is mainly identified for its efficiency, which is a result of total four optimization methods. The first of these methods is compressing both the original transactional database and the generated conditional databases into an FP-tree structure. FP-trees, also known as prefix trees, are constructed such that transactions with the same prefix share portions of the path down the tree. The details of this structure are complex; suffice to say that this effectively compresses the databases. The next two optimizations extract items in different ways.

The second extracts every item appearing in the conditional databases of the frequent item subsets. This helps reduction of the size of the FP-tree and improves the overall speed of the recursive process by combining some items. The third exploits the natural structure of the FP-tree by directly extracting frequent closed itemsets. Since the items have been arranged by prefix, this allows for natural harvesting of the closed itemsets. The final method prunes out frequent items which have the same level of support and can be expressed as a subset of other itemsets.

The CLOSET algorithm is a very efficient but highly complex technique. It allows for reasonably fast mining of frequent itemsets from data with control over the number of rules or itemsets generated. The mining procedure of CLOSET follows the FP-growth algorithm. However, the algorithm is able to extract only the closed patterns by careful book-keeping. CLOSET treats items appearing in every transaction of the conditional database specially. For example, if Q is the set of items that appears in every transaction of the P conditional database then P ∪ Q creates a frequent closed itemset if it is not a proper subset of any frequent closed itemset with the equal support. CLOSET also prunes the search space. For example, if P and Q are frequent itemset with the equal support where Q is also a closed itemset and P ⊂ Q, then it does not mine the conditional database of P because the latter will not produce any frequent closed itemsets [4].

```
Algorithm (CLOSET): Mining frequent closed itemsets by the FP-tree method
Input: Transaction dataset TDB and support threshold old min_sup;
Output: The complete set of frequent closed itemsets;
Method:
1. Initialization. Let FCI be the set of frequent closed itemset.
Initialize FCI□∅;
2. Find frequent items. Scan transaction database TDB, compute
frequent item list f-list;
3. Mine frequent closed itemsets recursively. Call CLOSET(∅,
TDB, f-list, FCI).
```

Sample Book Dataset Using Closet Algorithm

To prevent generating a huge number of book-sets, efficient mining frequent closed book-sets (CLOSET) is used. CLOSET is an FP-tree-based database projection method for efficient mining of frequent closed book-sets. This system explores the partition-based projection mechanism for scalable mining. It develops a single prefix path compression technique to identify frequent closed book-sets quickly [6].

For example, a book transactions database is considered. Firstly, books are named with code no.

Table 4.1 Book Information

CodeNo	BookName	Author
001	May I be love more	Juu
002	Still telling about love	Nanda TheinZan
003	Buddhism	U Nu
004	Min Lu's goodness	Min Lu
005	Flower for beautiful angel	Taryar Min Wai
006	Stupid person	Ni Ko Ye

Table 4.2 The book transactions database(TDB)

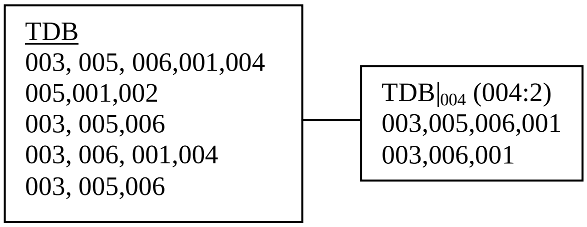
TransId	Books in Transaction
T1	003, 005, 006,001,004
T2	005,001,002
T3	003, 005,006
T4	003, 006, 001,004
T5	003, 005,006

Let minimum support count=2. In database, the set of frequent item list in support descending order: f_list={003:4, 005:4, 006:4, 001:3, 004:2}. By using Divide search space method, all frequent book-sets can be divided into 5 non-overlap subsets based on f_list:

- The ones containing 004.
- The ones containing 001 but no 004.
- The ones containing 006 but neither 001 nor 004.
- The ones containing 005 but neither 006 and 001 nor 004.

The ones containing only 003.

Then, the conditional database is constructed according to f_list to find closed frequent book-sets for each frequent item.



TDB
003, 005, 006,001,004
005,001,002
003, 005,006
003, 006, 001,004
003, 005,006

$TDB|_{004}$ (004:2)
003,005,006,001
003,006,001

Figure 4.1 Frequent closed book-sets for 004

From the above figure, F.C.I is {003, 006, 001, 004: 2}.

In the database, the system finds all transactions containing item 004. And the conditional database is constructed for 004, denoted as $TDB|_{004}$. In this database, only transactions containing 004 are included. But item 004 is omitted in each transaction since it appears in every transaction in the $TDB|_{004}$. And, all transactions containing in TDB|004 are all frequent. Then, the frequency of each item is counted. The support of 004 is 2. Items 003, 006 and 001 appear twice respectively in TDB|$_{004}$. This means that every transaction containing 004 also contains 003,006 and 001. Therefore, book-sets {003, 006, 001, 004: 2} is a frequent closed book-sets. When frequent book-sets containing 001 is searched, 004 have been found in TDB|$_{004}$ are omitted. And any items have been found are omitted in other transactions.

Generating Association Rules from Frequent Closed Book-Set

Once the frequent closed book-sets from transactions in a database D have been found, it is straightforward to generate strong association rules from them, where strong association rules satisfy both minimum support and minimum confidence. This can be done using the following equation:

Confidence is the measure of the strength of implication. Confidence (A=>B) = P(B|A) = support_count(A U B) / support_count(A). The conditional probability is expressed in terms of book-sets support count, where: support_count(AUB) is the number of transactions containing the book-sets AUB and support_count(A) is the number of transactions containing the book-sets A. Based on this equation, association rules can be generated as follows:

1. For each closed frequent book-sets l, generate all nonempty subsets of l.
2. For every nonempty subset s of l, output the rule “s => (l-s)” if support_count(l)/ support_count(s) >= min_conf, where min_conf is the minimum confidence threshold.

For example, a book transactions database shown in Figure 3.2 is considered. It issupposed that the data contain the closed frequent book-sets l = {003, 005, 006}. The non empty subsets of l are: {003, 005}, {003, 006}, 005, 006}, {003}, {005} and {006}. The resulting association rules are shown below, each listed with its confidence:

003 and 005=>006 Conf=3/3=100%

003 and 006=>005 Conf=3/4=75%

005 and 006=>003 Conf=3/3=100%

003=>005 and 006 (Conf=3/4=75%)

005=>003 and 006 (Conf=3/4=75%)

006=>003 and 005 (Conf=3/4=75%)

If the minimum confidence threshold is 80%, then only the first and third rules above are output, because these are the only ones generated that are strong [7].

V. ABOUT THE SYSTEM FLOW

According to the diagram in below, this system focuses on the association rule mining of data mining according to the related data of book borrowing shop. Firstly, data about the information of book borrowing shop is stored into the book database. In one transaction, book-sets of books borrowed by borrowers are contained. Firstly, transaction of books is stored in database. Then scan database and compute frequent item list, called f_list, defined by the user-specified minimum support count and sorted by descending frequency order.

Then, Divide search space method is used based on f_list. And conditional database is constructed from each frequent book-sets to find frequent closed book-sets. The pattern growth is achieved by the concentration of the suffix pattern with the frequent patterns generated from a conditional FP-tree. From the conditional FP-tree, the closed frequent book-sets using CLOSET algorithm is found. To facilitate tree traversal, a conditional database is built so that each book-set points to its occurrences in thetree via a chain of node-links. The tree with the associated node-links is obtained after scanning all of the transactions. In this way, the problem of book-sets in database is transformed to that of mining FP-tree.

Then association rules for these book-sets are generated. The confidence and correlation for the book-sets of books are calculated finally.

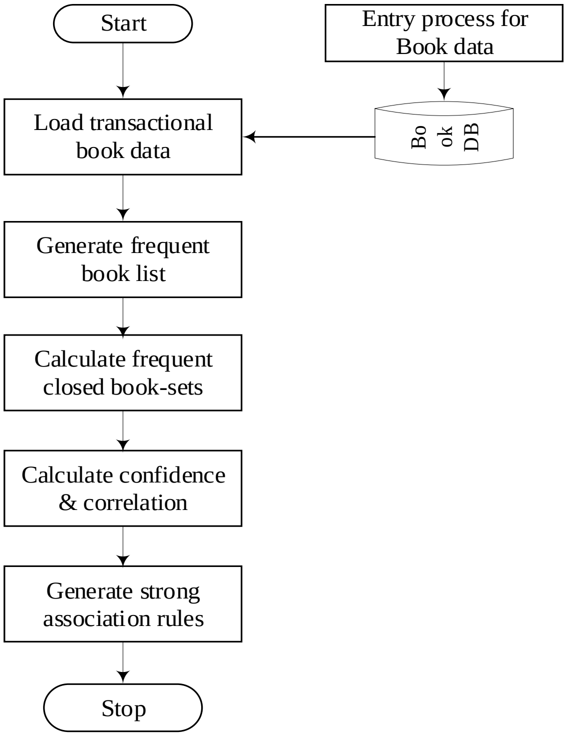

Start
Entry process for Book data
Load transactional book data
Bo ok DB
Generate frequent book list
Calculate frequent closed book-sets
Calculate confidence & correlation
Generate strong association rules
Stop

Figure 5.1 System Flow Diagram

Confidence for Frequent Closed Itemsets

From the above generated frequent closed book-sets of books, their associated information and their confidences are appeared. The set of frequent closed itemsets with their associated confidences are shown in Table 5.1.

Tabel 5.1 Associated Confidence of Booksets

F.C.I(A)	F.C.I(B)	Confi(%)	BookInfo
003	005	100	003= Buddhism(U Nu)(Dhamma Literature) 005=Flower for beauty angel(Taryar Min Wai)(Long story)
003	006	75	003=Buddhism(U Nu)(Dhamma Literature) 006=Stupid person(Ni Ko Ye)(Long story)
005	006	100	005=Flower for beauty angel(Taryar Min Wai)(Long story) 006=Stupid person(Ni Ko Ye)(Long story)
003	005,006	75	003=Buddhism(U Nu)(Dhamma Literature) 005=Flower for beauty angel(Taryar Min Wai)(Long story) 006=Stupid person(Ni Ko Ye)(Long story)

VI. RULE INTERESTINGNESS MEASURE BY CORRELATION ANALYSIS

A correlation measure can be used to augment the support-confidence framework for association rules. There are various correlations that measure to determine which would be good for mining large data sets. Lift is a simple correlation measure that is given as follows. The occurrence of book-sets A is independent of the occurrence of book-sets B if it is P(A∪B) = P(A)P(B); otherwise, book-sets A and B are dependent and correlated as events. This definition can easily be extended to more than two book-sets.

Although minimum support and confidence threshold help the exploration of a good number of uninteresting rules, many rules generated are still not interesting to the users. This is especially true when mining at low support threshold or mining for long patterns. Even strong interesting rules can be uninteresting or misleading. The support-confidence framework can be supplemented with additional interestingness measures based on correlation analysis.

VII. CONCLUSION

Frequent closed itemset mining discovers the complete set of frequent itemsets and the relationships between a given data set. The basic concept of frequent closed itemset mining is searching the interestingness and correlations between itemsets in transactional and relational database. To prevent generating a huge number of itemsets, efficient mining frequent closed itemsets(CLOSET) is used. CLOSET is apartitioning-based divide-and-conquer method for efficient mining of frequent closed itemsets. There are several advantages of CLOSET over other approaches:

- It constructs a highly compact FP-tree based on conditional databases.
- It avoids costly candidate generation and test by successively concatenating frequent 1-itemset found in the (conditional) FP-trees.
- It applies a partitioning-based divide-and-conquer method which dramatically reduces the size of the subsequent conditional pattern bases and conditional FP-tree.
- It develops a single prefix path compression technique to identify frequent closed itemsets quickly.

CLOSET algorithm which discovers interesting association or correlation relationships among huge number of data items reduce the redundant rules and find the complete set of itemsets that satisfy the minimum support item count. Mining complete set of itemsets suffers from generating a very large number of itemsets and association rules. Mining frequent closed itemsets provides an interesting alternative since it inherits the same analytical power as mining the whole set of frequent itemsets but generates a much smaller set of frequent itemsets and leads to less and more interesting rules than the former.

This system generates interesting and strong association rules. And correlation between strong association rules is calculated based on lift method. So, the user can evaluate that which book-sets are really interesting. By analysis of book categories, novel books are found that as the most borrowing books and thriller novels are the least borrowing books. Therefore, the user knows closed frequent book-sets of books and type of category using the CLOSET algorithm.

REFERENCES

[1]. Charu C. Aggarwal, Jiawei Han Editors, “Frequent Pattern Mining”.

[2]. C.I. Ezeife and Dan Zhang, “TidFP: Mining Frequent Patterns in Different Databases with Transaction ID”, School of Computer Science, University of Windsor, Windsor, Ontario, Canada N9B 3P4 zhang3d@uwindsor.ca, http://www.cs.uwindsor.ca/~cezeife.

[3]. David Hand, HeikkiMannila and Padhraic Smyth, “Principles of Data Mining” ISBN: 026208290xThe MIT Press © 2001 (546 pages)A comprehensive, highly technical look at the math and science behindextracting useful information from large databases.

[4]. DungarwalJayesh M and NeeruYadav, “A Review paper for mining Frequent Closed Itemsets”, S.V.C.S.E. Alwar Rajasthan – India and Prof S.V.C.S.E Alwar Rajasthan – India.

[5]. JagdishPanjwani and Mrs. ChitritaChaudhuri, “Application of FP Tree Growth Algorithm in Text Mining”, Project Report Submitted In Partial Fulfillment Of The Requirements for the Degree Of Master of Computer Application, Department of Computer Science and Engineering, Faculty of Engineering and Technology Jadavpur University, Kolkata-700032, India.

[6]. Jian Pen, Jiawei Han, and Runying Mao, “An efficient Algorithm for mining Frequent Closed Itemsets”,Intelligent Database Systems Research Lab, School of Computing Science, Simon Fraser University, Burnaby, B.C., CanadaV5A1S6, {peijian , han, rmao}@cs.sfu.ca.

[7]. Jiawei Han and MichelineKamber, “Frequent Item set Mining Methods, Data Mining– Concepts and Techniques”, Chapter 5.2, Julianna KatalinSipos.

[8]. Jiawei Han hanj@cs.uiuc.edu, “Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach”, University of Illinois at Urbana-Champaign.

[9]. KyaeHmon, "Applying Associative Rule Mining And Correlation Analysis On Software CD Selection Data", Computer University (Monywa), Myanmar, mirror315@gmail.com.

[10]. Lai Lai Win, KhinMyatMyat Moe, Computer University (Magway), "Mining Association Rules by using Vertical Data Format", lailaiwin.myn@gmail.com.

[11]. Mihir R Patel ,DipakDabhi, "An Extensive Survey on Association Rule Mining Algorithms", Assistant Professor, Department of Computer Engineering, CGPIT, Bardoli, India.

[12]. Springer, "Principle of Data Mining, Undergraduate Topics in Computer Science".

[13]. V. PurushothamaRaju and G.P. SaradhiVarma, "Mining closed sequential pattern in large database", Department of Information Technology S.R.K.R. Engineering College, Bhimavaram, A.P., India.