Efficient Implementation of Fp Growth Algorithm On Library Data

War War Myint[1], Hlaing Phyu Phyu Mon[2] & Hnin Yu Hlaing[3]
[1]University of Computer Studies (Meiktila), Faculty of Information Science
[2,3]University of Computer Studies (Meiktila), Faculty of Information Science

Abstract:

Data mining techniques are used in the field of many studies for various purposes. Everyday organizations collect huge amount of data from several resource. So, in this research, library data is considered as most famous application to mine that data to provide interesting patterns or rules for the future perspective. Implementation on it to generate rules and patterns using Frequent Pattern (FP)-Growth algorithm is the major concern of this research study. This study is to provide more guidance to the Liberian and the relation of a Liberian and a borrower. Frequent itemsets are generated based on the chosen borrowed books and minimum support value. The extracted frequent itemsets help the Liberian to make decisions which book is placed near at which book and determine the risk level of library data at an early stage. The proposed method can be applied to library dataset to predict the risk factors with risk level of the books based on chosen factors.

Key Words: Library Data, FP-Growth Algorithm

1. 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. 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 [7].

2. Problem Definition

Let I = { i1, i2, …. im} be the set of items and D be the transactional data source which contains the set of transactions. Each transaction T is a set of items such that T□I and is associated with an identifier called TID. An association rule is an implication of the form X=>Y, where X□I, Y□I and X_Y = _. In general, every association rule must satisfy two user specified constraints, one is support(_) and the other is confidence (_). The support of a rule X=>Y is defined as the fraction of transactions that contain X_Y, while the confidence is defined as the ratio of support(X_Y)/support(X). An itemset is frequent if its support satisfies at least the minimum support, otherwise it is said to be infrequent. A frequent itemset is a Maximal Frequent itemset if it is a frequent set and no superset of this is a frequent set. The paper aims to find the Maximal Frequent itemset from a huge data source.

3. Related Work

The solution is the frequent-pattern growth, or simply FP-growth, which mines the complete set of frequent itemsets without candidate generation. This method adopts a divide-and-conquer strategy as follows: first it compresses the database representing frequent items into frequent-pattern tree, or FP-tree, which retains the itemset association information. It then divides the compressed database into set of conditional databases; each associated with one frequent item or pattern fragment and mines each such database separately. FP-tree is created from the root and labels it null [9].

The FP-growth algorithm: (mine frequent itemsets using an FP-tree by pattern fragment growth):

Input:

1. D, a transaction database.
2. min_sup, the minimum support count threshold.

Output: the complete set of frequent patterns.

Method:

(1) The FP-tree is constructed.
(2) The FP-tree is mined by calling FP-growth (FP_tree, null):

Procedure FP_growth (Tree, α)

if Tree contains a single path P then
for each combination (denoted as β) of the nodes in the path P
generate pattern β U α with support_count = minimum support count of nodes in β;
else for each ai in the header of Tree{
generate pattern β=ai U α with support_count = ai.support_count
construct β's conditional pattern base and then β's conditional FP_tree Treeβ;
if Treeβ!= 0 then
callFP_growth(Treeβ, β); } [4].

Based on the above algorithm, association rules can be generated as follows:

2. For each frequent itemset l, generate all nonempty subsets of l.
3. 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.

Support and confidence are defined as:

Support (A -> B) = P (A∪ B)

Confidence (A->B) = P(A/B) [4,8].

Book-set Transaction

Let D be a database of transaction. Each transaction consists of a transaction identifier and a set of many books { LI0001, SE0002, DM0003, MZ0004, DS0005, …, EC7500, RP0550,……, 09500,…….} selected from the universe books of all possible descriptive book borrowed within one year. Table 1 shows the five items of book information as sample.

Table 3.1 Sample Book Information

BookID	BookName
LI100	Project Management
DS200	Ontology Engineering
CG300	Information Processing
MZ400	Design and Analysis
SE500	Java Programming

Table 3.2 Transaction of Book Borrowing

Transaction ID	BookName
T100	LI100, DS200,SE500
T200	DS200, MZ400
T300	DS200, CG300
T400	LI100, DS200, MZ400
T500	LI100, CG300
T600	DS200, CG300
T700	LI100, CG300
T800	LI100, DS200, CG300, SE500
T900	LI100, DS200, CG300

There are transactions of book borrowed in this database. In the process of mining frequent itemsets, the support count of an itemset is the length of the TID_set of the itemset. Suppose that the minimum transaction support count is 2.

BookId	Conditional Pattern Base	Conditional FP-tree	Frequent Pattern
SE500	{{DS200, LI100:1}, {DS200, LI100, CG300: 1}}	⟨DS200: 2, LI100: 2⟩	{DS200, SE500: 2}, {LI100, SE500: 2}, {DS200, LI100, SE500: 2}
MZ400	{{DS200, LI100: 1}, {DS200: 1}}	⟨DS200: 2⟩	{DS200, MZ400: 2}
CG300	{{DS200, LI100: 2}, {DS200, 2}, {LI100: 2}}	⟨DS200: 4, LI100: 2⟩, ⟨LI100: 2⟩	{DS200, CG300: 4}, {LI100, CG300: 4}, {DS200, LI100, CG300: 2}
LI100	{{DS200: 4}}	⟨DS200: 4⟩	{DS200, LI100: 4}

Figure 3.1 BookID-Frequency Tree

In figure 3.1, frequencies for each BookID are included after pruning with minimum support count 2. FP-Growth extracts frequent BookID from the FP-tree by using Bottom-up algorithm - from the leaves towards the root. It uses divide and conquer approach.

Divide and conquer:

- Compress the database (build FP-tree) to retain item-sets association information.
- Divides the compressed database into a set of conditional database.

Once the frequent itmesets from transaction in the database have been found, it is straightforward to generate association rules from them. This can be done using the following equation for the confidence, can be shown for completeness.

$$confidence\ (A \cup B) = \frac{support\_count(A \cup B)}{support(A)}$$

The resulting association rules are as shown below, each listed with its confidence:

LI100 ^ DS200 → SE500 (Confidence: 2/4 = 50%)

LI100 ^ SE500 → DS200 (Confidence: 2/2 = 100%)

DS200 ^ SE500 → LI100 (Confidence: 2/2 = 100%)

LI100 → DS200 ^ SE500 (Confidence: 2/6 = 33%)

DS200 → LI100 ^ SE500 (Confidence: 2/7 = 29%)

SE500 → LI100 ^ DS200 (Confidence: 2/2 = 100%)

International Journal of Research
Available at https://journals.pen2print.org/index.php/ijr/
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 06 Issue 07
June 2019

Result Set for Books and Its Category

Sup(A)	Sup(B)	Category
Project Management & Software Engineering	Java Programming	Software Engineering & Programming Language
Project Management & Java Programming	Software Engineering	Software Engineering & Programming Language
Software Engineering & Java Programming	Project Management	Software Engineering & Programming Language
Project Management	Software Engineering & Java Programming	Software Engineering & Programming Language
Software Engineering	Project Management & Java Programming	Software Engineering & Programming Language
Java Programming	Project Management & Software Engineering	Software Engineering & Programming Language

4. Design and Implementation

This system implemented for retrieving information of Book Borrowing by FP-Growth algorithm works the following procedure. In implementing this system, a database of book borrowed transactions is used. The database is used to send out learning. The database describes attributes of the books borrowed such as title, category, author_name publication_house, published_date, and so on. In this system, fp-growth algorithm is used together with 'divide and conquer approach'.

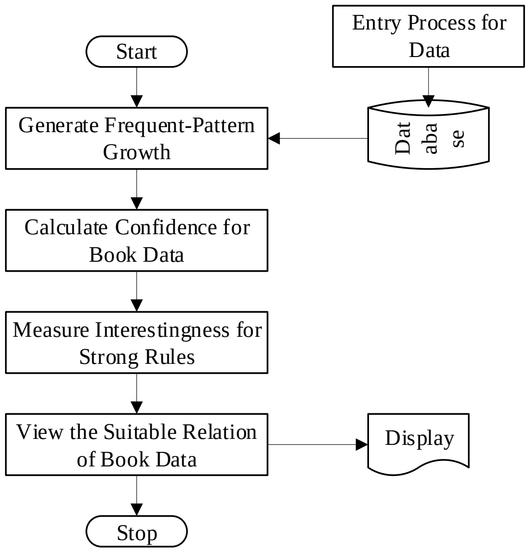

Figure 4.1 System Flow Diagram

This system focuses on the association rule mining of data mining according to the related data of books borrowed. Firstly, data about the information of Books borrowed is stored into the database. In one transaction, book_id occurred in borrowing are contained. By applying Frequent-Pattern growth algorithm of association rule mining, frequency of book_id are numbered with the minimum support count defined by the user-specified minimum support count and sorted by descending frequency order.

Implementation of the System: There are six steps to implement this system. This system works as follows.

- Entry process for data books borrowed and other related facts.
- FP-tree is constructed with book_id by the FP-growth algorithm.
- Generate frequent pattern of book_id.
- Calculate confidence for book borrowed.
- Search the suitable category of book_id to identify which types of book are more preferred by borrowers.
- Display the related (association rules) of books.

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 itemset A is independent of the occurrence of itemset B if P(A ∪B) = P(A)P(B); otherwise, itemsetsA and B are dependent and correlated as events. This definition can easily be extended to more than two itemsets. If the resulting value is less than 1, then the occurrence of A is negatively correlated with the occurrence of B. If the resulting value of a rule is greater than 1, then A and B are positively correlated, that is meaning that the occurrence of one implies the occurrence of the other. The lift between the occurrence of A and B can be measured by computing Lift (A, B) = P (A∪B)/ P (A)P(B).There are rule interestingness measures for above strong rules by lift as correlation analysis [13].

5. Conclusion

A novel data structure, frequent pattern tree (FP-tree), for storing compressed, crucial information about frequent patterns, and developed a pattern growth method, FP-growth, for efficient mining of frequent patterns in large databases are proposed. There are several advantages of FP-growth over other approaches: (1) It constructs a highly compact FP-tree. (2) It avoids costly candidate generation and test by successively concatenating frequent 1-itemset found in the (conditional) FP-trees. (3) It applies a partitioning-based divide-and-conquer method which dramatically reduces the size of the subsequent conditional pattern bases and conditional FP-tree.

Transactions of books are built by scanning from the database with FP-Growth Algorithm. Furthermore, other data related to this system such as can be stored. The association rules play a major role in many data mining application, trying to find interesting patterns in data bases. However, it is sometimes unrealistic to construct a main memory-based FP-tree.

Fp-Growth algorithm decomposes transaction records of books that is borrowed according to the frequent patterns obtained so far. It leads to focused search of smaller databases and compresses database called FP-tree structure.

6. Acknowledgements

I would like to take this opportunity to express my sincere thanks to all my seniors and associates who gave me a lot of valuable advice and information. I am also graceful to all respectable people who directly or indirectly contributed towards the success of this research. I especially also thank to my parents. I owe my respectful thanks to Dr. Mie Mie Khin, Rector of the University of Computer Studies, Meiktila for allowing me to develop this research and giving me general guidance during the period of studying time.

I owe a great debt of gratitude to Dr. Hlaing Phyu Phyu Mon, Professor and Head of Faculty of Information Science, University of Computer Studies, Meiktila, for giving advices.

7. Limitation and Further Extension

This system is implemented only for the related information of library data in UCSMTLA. The numbers of book transaction which are related to borrow to students or staffs in UCSMTLA can be added

In future, this system can be extended by adding many other transactions of books borrowed and can be improved by other association rule mining algorithms or frequent itemset mining algorithms.

REFERENCE

[1] Bhavesh V. Berani, Dr.ChiragThaker, Assistant Professors, “FP Growth Algorithm for finding patterns in Semantic Web”, Shantilal shah engineering college, Bhavnagar.

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

[3] 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.

[4]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.

[5] Jiawei Han and Micheline Kanber, “Data Mining Concepts and Technique, Second Edition”.

[6] Jiawei Han and Micheline Kamber, “Frequent Item set Mining Methods, Data Mining– Concepts and Techniques”, Chapter 5.2, Julianna Katalin Sipos.

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

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

[9] Springer, “Principle of Data Mining, Undergraduate Topics in Computer Science”.

[10] “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.

[11] “A Frequent Pattern Mining Algorithm Based on FP-growth without Generating Tree”, Universiti Putra Malaysia, Serdang, MALAYSIA, 1tohidi.h@gmail.com, 2hamidah@fsktm.upm.edu.my.

[12] “FP-Tree Based Algorithms Analysis, FPGrowth, COFI-Tree”, Thapar University, Patiala, India, bharatgupta35@gmail.com.

[13] Student of MSc (CS), Jrnrvu University, Udaipur, India, “Student’s Performance Prediction Using FP-Tree Data Mining Techniques”.

[14] M.A.Nishara Banu, B Gomathy, PG Scholar, Assistant Professor (Sr. G), Department of Computer Science and

Engineering, Bannari Amman Institute of Technology Sathyamangalam, India, "DISEASE PREDICTING SYSTEM USING DATA MINING TECHNIQUES".

[15] Ms. VibhavariChavan, Prof. Rajesh. N. Phursule, "Survey Paper On Big Data", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (6), 2014.

[16] Chun-Wei Tsai, Chin-Feng Lai, Han-Chieh Chao and Athanasios V. Vasilakos, "Big data analytics: A Survey", Journal of Big Data 2015.

[17] VarshaMashoria, Anju Singh, "Literature Survey on Various Frequent Pattern Mining Algorithm", IOSR Journal of Engineering (IOSRJEN), Vol. 3, Issue 1 (Jan. 2013), PP 58-64.

[18] VarshaMashoria, Anju Singh, "Literature Survey on Various Frequent Pattern Mining Algorithm", IOSR Journal of Engineering (IOSRJEN), Vol. 3, Issue 1 (Jan. 2013), PP 58-64.

[19] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A.I. Verkamo, "Fast discovery of association rules". In U.M., Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, "Advances in Knowledge Discovery and Data Mining", pages 307–328. MIT Press, 1996.

[20] Arpan Shah, Pratik A. Patel, "A Collaborative Approach of Frequent Item Set Mining: A Survey", International Journal of Computer Applications (0975 – 8887), Volume 107 – No 8, December 2014.

[21] S. Neelima, N. Satyanarayana and P. Krishna Murthy3, "A Survey on Approaches for Mining Frequent Itemsets", IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p-ISSN: 2278-87.

[22] ManishaGirotra, KanikaNagpal, SaloniMinocha, Neha Sharma, "Comparative Survey on Association Rule Mining Algorithms", International Journal of Computer Applications (0975 – 8887), Volume 84 – No 10, December 2013.

[23] Divya R., and Kumar, V.S. "Survey on AIS, Apriori and FP-Tree algorithm". International Journal of Computer Science and Management Research.Volume 1 Issue 2 September-2012 ISSN 2278-733X.

[24] K.R.Suneetha, R.Krishnamoorti, "Advanced Version of Apriori Algorithm", First International Conference on Integrated Intelligent Computing-2010, International Journal of Emerging Trends in Engineering and Development, Issue 5, Vol. 3 (April.-May. 2015).

[25] Xiaomei Yu, Hong Wang, "Improvement of Eclat Algorithm Based on Support in Frequent Itemset Mining", JOURNAL OF COMPUTERS, VOL. 9, NO. 9, SEPTEMBER 2014.