ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 9, September 2015

A Survey on Concise and Lossless Representation of Frequent Pattern Sets

R.Prabamanieswari[1], D.S.Mahendran[2], T.C. Raja Kumar[3]

Associate Professor, Computer Science, Govindammal Aditanar College for Women, Tiruchendur, India[1]

Associate Professor, Computer Science, Aditanar College of Arts & Science, Tiruchendur, India[2]

Associate Professor, Computer Science, St. Xavier's College, Tirunelveli, India[3]

Abstract: Many approaches are used to develop efficient algorithms for mining frequent patterns. Recent studies on frequent itemset mining algorithms resulted in significant performance improvements. However, if the minimum support is set to low, or the data is highly correlated, the number of frequent itemsets itself can be prohibitively large. To overcome this problem, several proposals have been made to construct a concise representation of the frequent itemsets, instead of mining all frequent itemsets. This survey paper illustrates the importance of FP-growth based algorithms for mining representative pattern sets. It also discusses that the number of representative pattern sets can be much smaller than the total number of frequent patterns; all the frequent patterns and their support can be recovered from the set of representative patterns. That is, the representative pattern sets are used to best approximate all frequent patterns.

Keywords: Depth-First strategy, FP- growth, frequent itemset**,** closed itemset, Representative Pattern set.

I. INTRODUCTION

Data Mining is the process of extracting previously unknown and potentially useful hidden predictive information from large amounts of data. In general, data mining tasks can be classified into two categories such as descriptive and predictive. Descriptive mining tasks characterize the general properties of the data in the database. Predictive mining tasks perform inference on the current data in order to make predictions. One of the descriptive mining, Association Rule Mining plays an important role in finding frequent itemsets. The Association Rule Mining has two steps namely frequent itemset mining and association rule generation. Furthermore, frequent itemsets play an essential role in many data mining tasks that try to find interesting patterns from databases, such as correlations, sequences, episodes, classifiers, clusters etc. A frequent itemset typically refers to a set of items that frequently appear together in a transactional data set, such as milk and bread. Frequent pattern mining can be classified in various ways such as based on the completeness of patterns to be mined, based on the levels of abstraction involved in the rule set, based on the number of data dimensions involved in the rule, based on the types of values handled in the rule, based on kinds of rules to be mined and based on the kinds of patterns to be mined.This paper concentrates the method which is based on the kinds of patterns to be mined such as representative patterns.

Many approaches are used to develop efficient algorithms for mining frequent patterns. Various search strategies have been developed, such as depth-first search vs. breadth-first search, vertical formats vs. horizontal formats, tree-structure vs. other data structures, top-down vs. bottom-up traversal, pseudo projection vs. physical projection of conditional database, etc. They can be classified as Apriori [1] based algorithms and FP-growth based algorithms [3].The FP-growth method, which explores some compressed data structure such as FP-tree. The FP-tree is a compact representation of all relevant frequency information in a database. Compression is achieved by building the tree in such a way that overlapping itemsets share prefixes of the corresponding branches. The FP-tree has a header table associated with it. Single items and their counts are stored in the header table in decreasing order of their frequency. The entry for an item also contains the head of a list that links all the corresponding nodes of the FP-tree. Compared with Apriori [1] and its variants which need several database scans [2], the FP-growth method [3] only needs two database scans when mining all frequent itemsets. The first scan counts the number of occurrences of each item. The second scan constructs the initial FP-tree which contains all frequency information of the original dataset. Mining the database then becomes mining the FP-tree. The FP-tree can be searched by following the depth-first strategy. The method FP-growth is about an order of magnitude faster than the Apriori.

In the early days, the size of the database and the generation of a reasonable amount of frequent itemsets were considered as the most costly aspects of frequent itemset mining, and the most energy went into minimizing the number of scans through the database. However, if the minimal support threshold is set too low, or the data is highly correlated, the number of frequent itemsets itself can be prohibitively large. To overcome this problem, recently several proposals have been made to construct a concise representation based on lossless compression methods such as closed itemsets [4-9] and constraints based frequent itemsets [10-13] instead of mining all

DOI 10.17148/IJARCCE.2015.4927

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 9, September 2015

frequent itemsets. The constraint based mining though useful, but can hardly be used for pre-computation, since different users are likely to have different constraints. The concise representation (closed itemsets) gives the less number of frequent representative patterns and all the frequent itemsets can be derived from them with exact support value. But, most of the applications will not need precise support information for frequent patterns: a good approximation for the support count could be more than adequate. Here, by a good approximation, we mean that the frequency of every frequent pattern can be estimated with a guaranteed maximal error bound. Therefore, the main goal of this paper is to point up the previously used approaches which are based on FP-growth algorithm for mining representative patterns to reduce the number of frequent itemsets.

The rest of the paper is organized as follows: Section II gives the basic definitions of frequent itemsets, frequent closed itemsets and representative pattern sets. Section III describes the related work. Section IV summarizes the approaches which are based on depth-first strategy and lossless compression method. Finally, Section V concludes the paper.

II. BASIC DEFINITIONS

This section gives the basic definitions of frequent itemsets, frequent closed itemsets and representative pattern sets.Let $I=I_1, I_2, \dots, I_m$ be a set of m distinct attributes, T be transaction that contains a set of items such that $T \subseteq I$ and D be a database with different transaction records Ts.

Definition 1. (Items and itemsets).
An itemset X is a finite subset of I, the set of possible items.

Definition 2. (Transactions and transaction databases).
A transaction t is a pair $\prec i,X\succ$ consisting of a transaction identifier tid (t) = i $\in$ N and an itemset (t) = X $\subseteq$ I. A transaction database D is a set of transactions with unique transaction identifiers. The set S_D of all itemsets in D is $S_D = \{X: \prec i, X\succ i \in D\}$.

Definition 3. (Support/frequencies).
The support of X in D is supp(X, D) = |cover(X, D)| where cover (X, D)= {tid | (tid, I) ϵ D, X $\subseteq$I}.

Definition 4. (Frequent itemset mining).
Given a transaction database D and a real value $\sigma \in [0, 1]$, find all σ frequent itemsets, i.e., determine the collection F (σ, D) = {X$\subseteq$I: support (X, D) $\geq$ σ} of σ - frequent itemsets in D.

Definition 5. (Closed frequent itemsets).
An itemset $X \in F(\sigma, D)$ is closed, if there exists no proper super-itemset Y such that Y has the same support count as X in F. The collection of closed σ -frequent itemsets in D is denoted by C (σ, D).

Definition 6. (Representative pattern sets). Given a real number $\epsilon[0, 1]$ and two patterns $X_1$and X_2, we say $X_1$is ϵ-covered by X_2 if $X_1 \subseteq X_2$and $D(X_1, X_2) \leq \epsilon$ where $D(X1, X2) = 1 - \frac{|T(X1\cup)\cap T(X2)|}{|T(X1)\cup T(X2)|}$. The Pattern X_2 approximates the pattern$X_1$and we say that $X_2$is a representative pattern set.
If we use supp(X_2) to approximatesupp(X_1), then the relative error supp(X_1) $-$supp(X_2)/supp(X_1) is no largerthan ϵ.

III.RELATED WORK

Many algorithms and techniques are proposed for enumerating itemsets from transactional databases. It has been observed that the complete set of frequent patterns often contains a lot of redundancy i.e.) many frequent patterns have similar items and supporting transactions. To overcome this problem, several approaches have been made to construct a concise representation of the frequent itemsets. Two major approaches have been developed in this direction: lossless compression and lossy approximation. The closed frequent patterns [5] and non-derivable itemsets [17] methods are generally referred to as lossless compression since we can fully recover the exact frequency of any frequent itemsets. The maximal frequent pattern [21] is called as lossy compression since we cannot recover the exact frequencies. In addition to these approaches, recently many proposals such as generators [18], disjunction-free generators [19], δ-free sets [20], top-k frequent closed patterns [22] and redundancy-aware top k patterns [23] have been made to construct a concise (compressed) representation of the frequent itemsets, instead of mining all frequent itemsets. But, the type of concise representation that received a lot of attention in the literature is the closed itemsets because the number of closed patterns is always smaller or equal in cardinality than the set of frequent free sets [19-20], lesser than that of generators [18] and the number of non-derivable patterns [17] is larger than that of closed patterns on some datasets. Furthermore, the set of generators itself is not lossless. Hence, this study focuses to find a minimum representative pattern set based on the lossless compression method such as closed itemset.

A. Lossless Compression Approach

Frequent closed patterns preserve the exact support of all frequent patterns. The concept of closed frequent patterns is proposed by Pasquier et al. The Close [4-5] and the A-Close [6] algorithms perform the breadth first search for the generators of the frequent closed itemsets in a level wise manner. These kinds of patterns are concise in the sense that all of the frequent patterns can be derived from them. Unfortunately, the number of patterns generated in these approaches is still too large to handle. CLOSET [8] is an extension of the FP-growth algorithm [3] which constructs a frequent pattern tree FP-tree and recursively builds conditional FP-trees in a bottom-up tree search manner. Although CLOSET uses several optimization techniques to enhance the mining performance, its performance still suffers in sparse datasets or when the support threshold is low. The algorithm CLOSET+ [9] uses one global prefix-tree for keeping track of all closed itemsets. FP close [16] is one of the best algorithms for

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 9, September 2015

mining closed frequent itemsets, even when compared to CLOSET+. In this algorithm, a CFI-tree, another variation of the FP-tree, is used for testing the closeness of frequent itemsets. But, these frequent closed patterns group the patterns supported by exactly the same set of transactions together. This condition is too restrictive. Therefore, the constraints can be used to capture the users' focus, and effective strategies have been developed to push various constraints deep into the mining process [10-13]. Even though these approaches are useful, they may still be insufficient in some situations.

Compression using the closed-pattern approach may not be very effective, since slightly different counts often exist between super and sub patterns. Constraint based mining, though useful, can hardly be used for pre-computation, since different users are likely to have different constraints. Most applications will not need precise support information for frequent patterns: a good approximation for the support count could be more than adequate. Here, by a good approximation, we mean that the frequency of every frequent pattern can be estimated with a guaranteed maximal error bound.

B. Representative Pattern Set Approach

Previously, the ideas of approximating frequent patterns have been probed in some related studies. For example, Mannila and Toivonen [14] show that approximate association rules are interesting and useful. In [20], the notion of free-sets is proposed and it can be used to approximate closely the support of frequent itemsets. However, none of these studies systematically explored the problem of designing and mining condensed frequent-pattern bases with a guaranteed maximal error bound. Disjunction-free generators [19] and δ-free sets [20] give the support of all frequent patterns approximately. Both disjunction-free generators and δ-free sets also require a border to be lossless. Jian Pei et al [15] consider two types of condensed FP-bases: the downward condensed FP-base Bd and the max-pattern-based condensed FP-base Bm. They also specify that computing a condensed FP-base can also be performed on a relative, percentage based error bound k% instead of an absolute error bound k. In that case, supub−suplb / suplb≤ k% should be satisfied for frequent patterns. But, all these approaches are less efficient than the approach which is based on closed pattern approach to approximate the support count.

Recently, several approaches have been proposed to tackle the concise representation of frequent itemsets, two key criteria being employed for evaluating the concise representation of itemsets are the coverage criterion and frequency criterion. The methods like top-k frequent patterns [22], top-k redundancy-aware patterns [23], and error-tolerant patterns [24]try to rank the importance of individual patterns, or revise the frequency concept to reduce the number of frequent patterns. But, choosing an appropriate k for a given domain is usually not easy and there are no theoretical guarantees on the level of approximation for a given k. The major problem with these approaches is that the frequency (or the support measure) is not considered. However, these methods generally do not provide a good representation of the collection of frequent patterns. Therefore, this study concentrates both the key criteria for the concise representation of itemsets.

Xin et al. [25] propose the concept ofδ-covered to generalize the concept of frequent closed pattern. A pattern X1 is δ-covered by another pattern X_2 if $X_1$is a subset of X_2 and $(\text{supp}(X_1) - \text{supp}(X_2))/\text{supp}(X_1) \leq \delta$. They develop two algorithms, RPglobal and RPlocal. These algorithms need to perform substantial coverage checking that checks whether an item set can be covered by another one. RPglobal is very time-consuming and space-consuming. It is feasible only when the number of frequent patterns is not large. RPlocal is very efficient, but it produces more representative patterns than RPglobal. To improve the performance, RPglobal and RPlocal have to use some FP-tree-like structures to index frequent item sets and representative item sets to reduce the number and the cost of coverage checking.

Jianzhong Li et al. [26] devise two algorithms, RP-FP and RP-GD, to mine a representative set that summarizes frequent sub graphs. RP-FP derives a representative set from frequent closed sub graphs, whereas RP-GD mines a representative set from graph databases directly. Based on the concept of δ –cover, they proposed three new concepts like jump value, δ -jump pattern, and δ -cover graph but these concepts are only used for graph mining.

Liu et al [27] analyse the bottlenecks of RPglobal and RPlocal and develop two algorithms, MinRPset and FlexRPset, to solve the problem. The algorithm MinRPset is similar to RPglobal, but it utilizes several techniques to reduce running time and memory usage. In particular, MinRPset uses a tree structure called CFP-tree [28] to store frequent patterns compactly. The algorithm FlexRPset is developed based on MinRPset. It provides one extra parameter K, which allows users to make a trade-off between efficiency and the number of representative patterns selected. In [25] and [27], the relative error $(\text{supp}(X_1) - \text{supp}(X_2))/\text{supp}(X_1)$ is used. But, MinRPset has the extra benefits besides giving fewer representative patterns.

IV. SUMMARY

Discovering Frequent Itemset is considered to an important research oriented task in data mining, due to its large applicability in real world applications. In recent years, many algorithms and techniques are proposed for enumerating itemsets from transactional databases. Apriori [1] is a bottom-up, breadth-first search Algorithm which uses monotonicity property: all supersets of an infrequent itemset must be infrequent. It enforces several scans through the database. Therefore, this study concentrates FP-growth algorithm. This section summarizes the various types of mining approaches such as frequent itemsets, frequent closed itemsets and representative pattern sets which are based on depth-first strategy and lossless compression method.

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 9, September 2015

TABLE I Depth-First Search Algorithms

TYPE OF MINING	ALGORITHM	PROCEDURE USED
Frequent Itemset Mining	FP-growth	FP-tree, Depth-First Search Two scanning of database Without Candidate Generation
Closed Itemset Mining (Lossless Compression)	CLOSET CLOSET+ FP close.	FP-tree, Depth-First Search FP-tree, Depth-First Search CFI-tree, Depth-First Search
Representative Pattern Set Mining (Approximation Approach)	RPglobal and RPlocal MinRPset and FlexRPset	RP-tree, Depth-First Search CFP-tree, Depth-First Search

V. CONCLUSION

A number of proposals have been made to construct a concise and lossless representation of frequent patterns such as closed frequent patterns, and non-derivable frequent itemsets. These kinds of patterns are concise in the sense that all the frequent patterns can be derived from them with exact support value. Unfortunately, the number of patterns generated in these two approaches is still too large to handle. Therefore, the research focuses the lossless compression methods to summarize the frequent patterns with a guaranteed error bound. This paper first presents the importance of FP-growth algorithm for mining frequent itemsets and then point ups the previously used approaches for reducing the number of frequent itemsets based on concise (compressed) representation such as closed frequent itemsets. Finally, the ideas of approximating frequent patterns from representative sets have been discussed based on depth-first search strategy.

REFERENCES

[1] R. Agrawal, T. Imielinski, and A. N. Swami, "Mining association rules between sets of items in large databases", in Proc. SIGMOD, Washington, DC, USA, pp. 207–216, 1993.
[2] Borgelt C, "Efficient implementations of apriori and éclat", In: Proceedings of the IEEE ICDM Workshop on Frequent Itemset Mining Implementations (FIMI'03). Volume 90 of CEUR Workshop Proceedings, Melbourne, Florida, USA, 2003.
[3] J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation", In Proceedings of ACM SIGMOD'00, pp 1–12, May 2000.
[4] Pasquier N, Bastide Y, Taouil R and Lakhal, "Pruning Closed Itemset Lattices for Association Rules", Proc. BDA conf., pp 177-196, 1998.
[5] Pasquier N, Bastide Y, Taouil R and Lakhal L, "Efficient Mining of Association Rules using Closed Itemset Lattices", Information Systems, vol 24, No 1, pp 25-46,1999.
[6] Pasquier N, Bastide Y, Taouil R and Lakhal, " Discovering frequent closed itemsets for association rules", In: Proc 7th Int Conf on Database Theory (ICDT'99), Jerusalem, Israel, pp 398–416,1999.
[7] Zaki M, "Generating non-redundant association rules", In: Proc 2000 ACM SIGKDD Int Conf KnowledgeDiscovery in Databases (KDD'00), Boston, USA, pp 34–43, 2000.
[8] J. Pei, J. Han, and R. Mao, "CLOSET: An efficientalgorithm for mining frequent closed itemsets",In ACM SIGMOD'00 Workshop on Research Issues in Data Mining and KnowledgeDiscovery, pp 21–30, 2000.
[9] J. Wang, J. Han and J. Pei, "Closet+: Searching for the best strategiesfor mining frequent closed itemsets," in Proc. KDD, New York, NY, USA, pp. 236–245, 2003.
[10] Ng R, Lakshmanan LVS, Han J and Pang A, " Exploratory mining and pruning optimizations of constrainedassociations rules", In: Proc 1998 ACM-SIGMOD Int Conf Management of Data (SIGMOD'98), Seattle, USA, pp 13–24,1998.
[11] Lakshmanan LVS, Ng R, Han J and Pang A, " Optimization of constrained frequent set queries with2-variable constraints", In: Proc 1999 ACM-SIGMOD Int Conf Management of Data (SIGMOD'99),Philadelphia, USA, pp 157–168,1999.
[12] Pei J, Han J and Lakshmanan LVS, "Mining frequent itemsets with convertible constraints", In: Proc 2001Int Conf Data Engineering (ICDE'01), Heidelberg, Germany, pp 433–332, 2001.
[13] R. Srikant, Q. Vu, R. Agrawal, Mining association ruleswith item constraints, in: Proceedings of the 3rd ACMSIGKDD Conference, 1997, pp. 67–73.
[14] Mannila H and Toivonen H, "Multiple uses of frequent sets and condensed representations (extended abstract)", In: Knowledge Discovery and Data Mining, pp 189–194,1996.
[15] Jian Pei, Guozhu Dong, Wei ZouandJiawei Han,"Mining Condensed Frequent-Pattern Bases", Knowledge and Information Systems, 2004, DOI 10.1007/s10115-003-0133-6.
[16] Gosta Grahne and Jianfei Zhu, "Fast Algorithms for FrequentItemset Mining Using FP-Trees", IEEE Transactions on Knowledge and Data Engineering, vol. 17, no. 10, October 2005.
[17] T. Calders and B. Goethals., "Mining all non-derivablefrequent itemsets", In Proc. of 2002 European Conf. onPrinciples of Data Mining and Knowledge Discovery(PKDD'02), pp 74–85, 2002.
[18] Y. Bastide, N. Pasquier, R. Taouil, G. Stumme, and L. Lakhal,"Mining minimal non-redundant association rules using frequent closed itemsets," in Proc. 1st Int. Conf. CL, London, U.K., pp. 972–986,2000.
[19] A. Bykowski and C. Rigotti, "A condensed representation to find frequent patterns," in Proc. PODS, New York, NY, USA, pp. 267-273, 2001.
[20] J.F. Boulicaut, A. Bykowski, and C. Rigotti, "Free-sets: A condensed representation of boolean data for the approximation of frequency queries," Data Mining Knowl. Discov., vol. 7, no. 1, pp.5–22, 2003
[21] R. J. Bayardo, "Efficiently mining long patterns from databases, "in Proc. SIGMOD, New York, NY, USA, pp. 85–93, 1998.
[22] J. Wang, J. Han, Y. Lu, and P. Tzvetkov, "TFP: An efficient algorithm for mining top-k frequent closed itemsets," IEEE Trans. Knowl. Data Eng., vol. 17, no. 5, pp. 652–664, May 2005.
[23] D. Xin, H. Cheng, X. Yan, and J. Han, "Extracting redundancy-awaretop-k patterns," in Proc. KDD, Philadelphia, PA, USA, pp. 444–453,2006.
[24] M. T. Yang, R. Kasturi, and A. Sivasubramaniam., "An Automatic Scheduler for Real-Time Vision Applications", InP roceedings of the International Parallel and Distributed Processing Symposium(IPDPS), 2001.
[25] D. Xin, J. Han, X. Yan, and H. Cheng, "Mining compressed requent-pattern sets," in Proc. 31st Int. Conf. VLDB, Trondheim, Norway, pp. 709–720,2005.
[26] Jianzhong Li, Yong Liu, and Hong Gao, Efficient Algorithms for Summarizing Graph Patterns, IEEE Transactions on Knowledge and Data Engineering, Vol. 23, No. 9,pp 1388-1405, 2011.
[27] Guimei Liu, Haojun Zhang, and Limsoon Wong, "A Flexible Approach to Finding Representative Pattern Sets", IEEE Transactions on Knowledge and Data Engineering, Vol. 26, No. 7,pp 1562-1574,, July 2014.
[28] G. Liu, H. Lu, and J. X. Yu, "CFP-tree: A compact disk-based structure for storing and querying frequent itemsets" ,Inf. Syst., vol. 32, no.2, pp. 295–319, 2007.