Binary Journal of Data Mining & Networking 5 (2015) 26-29

http://www.arjournals.org/index.php/bjdmn/index

Research Article

ISSN: 2229 -7170

FP-Growth Tree Based Algorithms Analysis: CP-Tree and K Map

Neelesh Shrivastava[1] , Richa Khanna[2]

*Corresponding author:

Neelesh Shrivastava

[1]Dept of Computer Engineering V.I.T.S, Satna (M.P)
[2]Dept of CSE (Software System) V.I.T.S, Satna (M.P)

Abstract

We propose a novel frequent-pattern tree (FP-tree) structure; our performance study shows that the FP-growth method is efficient and scalable for mining both long and short frequent patterns, and is about an order of magnitude faster than the Apriori algorithm and also faster than some recently reported new frequent-pattern mining methods. FP-tree method is efficient algorithm in association mining to mine frequent patterns in data mining, in spite of long or short frequent data patterns. By using compact best tree structure and partitioning-based and divide-and-conquer data mining searching method, it can be reduces the costs search substantially .it just as the analysis multi-CPU or reduce computer memory to solve problem. But this approach can be apparently decrease the costs for exchanging and combining control information and the algorithm complexity is also greatly decreased, solve this problem efficiently. Even if main adopting multi-CPU technique, raising the requirement is basically hardware, best performance improvement is still to be limited. Is there any other way that most one may it can reduce these costs in FP-tree construction, performance best improvement is still limited.

Keywords: partitioning-based, parallel, Projection, data mining, AI, Information.

Introduction

Efficiency of mining is achieved with three techniques: Our FP-tree-based mining adopts a pattern-fragment growth method to avoid the costly generation of a large number of candidate sets. Divide-and-conquer method is used to decompose the mining task into a set of smaller tasks for mining confined patterns in conditional databases, which dramatically reduces the search space Frequent pattern mining and association rule mining Lower call this temp database as Projection Database, we can create a temp database for storing all the frequent items ordered by the list of frequent items which is used for projecting,[1][3] reduce the expensive costs of individual node computation The case that may happen in a very large database.

We examine the size of FP-tree as well as the turning point of FP-growth on data projection to building FP-tree.

(1) The shared parts can be merged using one prefix structure as long as the count is registered properly.

(2)If two transactions share a common prefix, according to some sorted order of frequent items.

FP-TREE

Given a transaction database DB and a support threshold ξ there is several important properties in FP-tree that can be derived from the FP-tree construction process.

Frequent Patterns -Tree

Construction of a compact FP-tree ensures that subsequent mining can be performed with a rather compact data structure. This does not automatically guarantee that it will be highly most efficient since one may still encounter the combinatorial problem of candidate generation if one simply uses this FP-tree to generate and check all the candidate patterns [5,6,7].

Literature Review

There are number of different approaches available in the literature for frequent pattern mining from uncertain data [1], [10], [11], [13], [14], [20], [21]. This section provides some background and discusses work related to data uncertainty. Some researchers have extended association rule mining techniques to imprecise or uncertain data. They have proposed different approaches and framework.

Leung, et. al. proposed efficient algorithms for the mining of constrained frequent patterns from uncertain data [8] in 2009. They proposed, using U-FPS algorithms, to find the frequent patterns for efficient mining that satisfy the user-specified constraints from uncertain data.

Aggarwal, et. al. proposed a framework for clustering uncertain data streams [9] in 2008. They provide a method for clustering. They use a general model of the uncertainty in which they assume that only a few statistical measures of the uncertainty are available.

Chui, et. al. proposed mining a frequent item set from uncertain data [10] in 2007. They proposed the U-Apriori algorithm, which was a modified version of the Apriori algorithm, which works on such datasets. They identified the computational problem of U-Apriori and proposed a data trimming framework to address this

issue. They proposed a framework for mining frequent item sets from uncertain data. A data trimming framework proposed to improve mining efficiency. Through extensive experiments, the data trimming technique can achieve significant savings in both CPU cost and I/O cost.

Aggrawal, et. al. proposed frequent pattern mining with uncertain data [11] in 2009. They proposed several classical mining algorithms for deterministic data sets, and evaluated their performance in terms of memory usage and efficiency. The uncertain case has quite different trade-offs from the deterministic case because of the inclusion of probability information.

Abd-Elmegid, et. al. proposed vertical mining of frequent patterns from uncertain data [13] in 2011. They extended the state-of-art vertical mining algorithm, Eclat, for mining frequent patterns from uncertain data and generated the Eclat algorithm. In this paper they studied the problem of mining frequent itemsets from existential uncertain data using the Tid set vertical data representation. They also performed a comparative study between the proposed algorithm and well known algorithms.

Tang, et. al. proposed mining probabilistic frequent closed item sets in uncertain databases [14] in 2011. In this paper they pioneer in defining probabilistic frequent closed item sets in uncertain data. They proposed a probabilistic frequent closed item set mining (PFCIM) algorithm to mine from uncertain databases.

Ngai, et. al. proposed efficient clustering of uncertain data [22] in 2006. In this paper they studied the problem of uncertain object with the uncertainty regions defined by pdfs. They describe the min-max-dist pruning method and showed that it was fairly effective in pruning expected distance computations. They used four pruning methods, which was independent of each other and can be combined to achieve an even higher pruning effectiveness.

Leung, et. al. proposed the efficient mining of frequent patterns from uncertain data [23] in 2007. In this paper they proposed a tree-based mining algorithm (UFP-growth) to efficiently find frequent patterns from uncertain data, where each item in the transactions is associated with an existential probability. They plan to investigate ways to further reduce the tree size.

We briefly describe our basic approach to the problem and then produce the best results. In this paper, uncertain textual data is used to generate the frequent patterns.

FP Tree, CP-Tree and K Map

FPtree: Construction of a compact FP-tree ensures that subsequent mining can be performed with a rather compact data structure. However, this does not automatically guarantee that it will be highly efficient since one may still encounter the combinatorial problem of candidate generation if one simply uses this FP-tree to generate and check all the candidate patterns. In this section, we study how to explore the compact information stored in an FP-tree, develop the principles of frequent-pattern growth by examination of our running example, explore how to perform further optimization when there exist a single prefix path in an FP-tree, and propose a frequent-pattern growth algorithm, *FP-growth*, for mining the *complete set of frequent patterns* using FP-tree.

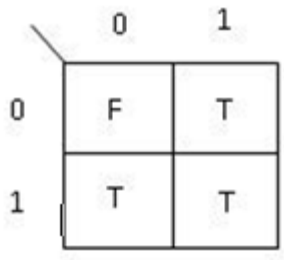
CP tree: First scan the database and manage the items appearing in the transaction. Then all the items whose support is less than the minimum support which is user defined are considered as infrequent are deleted from Consideration. All other remaining items are considered as frequent items and arrange in the sorted order of their frequency. This list is known as header table when store in table. All the respective support of the items is stored using pointers in the frequent pattern tree. Then construct the frequent pattern tree which is also known as compact tree. The sorted items according to frequency in header table are used to build the FP-tree. This needs complete database scan. when the item insert in the tree checks if it exist earlier in tree as in same order then increment the counter of support by one which is mentioned along with each item in the tree separated by comma, otherwise add new node with 1 as a support counter. A link is maintained using pointers which same item and its entry in header table. In header table, pointer points to the first occurrence of each item.

K Map: A Karnaugh map [18], [19] provides a pictorial method of grouping together expressions sharing common factors thus eliminating unrelated variables. A karnaugh map reduces the need for extensive calculation by taking advantage of the humans´ pattern recognition capability.

Result

This also permits the rapid identification and elimination of potential race conditions. A Karnaugh map is composed of many grid boxes. Each grid box in a k-map corresponds to a min term or max term. Using the defined min terms, the truth table can be created as a two variables in Table 1 and Figure 1.

1	1	T
1	0	T
0	1	T
0	0	F



Variables in k-map, Figure 1: General case of a two.

If the number of terms n is even then matrix of size 2n=2 2n=2 is created and if the number of terms n is odd then a matrix of size2 (n 1)=2 2(n 1)=2 created. In this research study, the k-map approach on uncertain textual data to find a frequent term set which reduces the database scans and improves the efficiency and accuracy of algorithm.

Table 1: Truth table for two variables

Algorithm Parameter	FP-Growth	CP-Tree	K Map
Structure	Simple Tree Based Structure	Uses Bidirectional FP-Tree Structure.	Uses compressed FP-Tree data Structure.
Approach	Recursive	Non- Recursive	Non- Recursive
Technique	It constructs conditional frequent pattern tree and conditional pattern base from database which satisfy the minimum Support.	It constructs bidirectional FP-Tree and builds the CP-tree-Trees for each item then mines the CP-Tree locally For each item.	It constructs the compact FP-Tree through mapping into index and then mine frequent item sets according to projections index separately
Memory Utilization	Low as for large database complete Tree structure cannot fit into main memory	Better, Fit into main memory due to mining locally in parts for the complete tree, Thus every part represent in main memory	Best, as Compress FP-tree structure used and mine according to projections separately thus easily fit into main memory
Databases	Good for dense databases	Good for dense as well as Sparse Databases. But with low support in sparse databases performance Degrades.	Good for dense as well as for Sparse databases.

Conclusion

FP-Growth is the first successful tree base algorithm for mining the frequent item sets. As for large databases its structure does not fit into main memory therefore new techniques come into pictures which are the variations of the classic FP-Tree. FP-Growth recursively mine the frequent item sets but some variations CP - Tree and K Map based upon non recursive. Pruning method consists of removing all the locally non frequent items and also CP-Tree and K map need less memory space and comparatively fast in execution then the FP-Tree because of their compact and different mining techniques. We want to improve efficiency in FP tree so apply Parallel and partition technique but both techniques are based on projection.

References

[1]. Aggarwal C C. An Introduction to uncertain data algorithm and applications, Advances in Database Systems. 2009; 35; 1–8.

[2]. Rajput D S, Thakur R S, Thakur G S. Rule Generation from Textual Data by using Graph Based Approach, International Journal of Computer Application (IJCA). 2011; 31(9); 36–43.

[3]. Han I, Kamber M. Data Mining concepts and Techniques, M. K. Publishers. 2000; 335–389.

[4]. Rajput D S, Thakur R S, Thakur G S. Fuzzy Association Rule Mining based Frequent Pattern Extraction from Uncertain Data, IEEE 2nd World Congress on Information and Communication Technologies (WICT'12). 2012; 709–714.

[5]. Thakur R S, Jain R C, Pardasani K R. Graph Theoretic Based Algorithm for Mining Frequent Patterns, IEEE World Congress on Computational Intelligence Hong Kong. 2008; 629–633.

[6]. Agrawal R, Srikant R. titFast algorithms for mining association rules In Proc. VLDB 1994, pp.487–499.

[7]. Rajput D S, Thakur R S, Thakur G S. Fuzzy Association Rule Mining based Knowledge Extraction in Large Textual Dataset, International Conference on Computer Engineering Mathematical Sciences (ICCEMS'12). 2012; 191–194.

[8]. Leung C K S, Hao B. Efficient algorithms for mining constrained frequent patterns from uncertain data, Proceedings of the 1st ACM SIGKDD Workshop on Knowledge Discovery from Uncertain Data. 2009; 9-18.

[9]. Aggarwal C C, Philip S Yu. A Framework for Clustering Uncertain

Data Streams, Data Engineering, IEEE 24th International Conference on ICDE'08. 2008; 150-159.

[10]. Chui C K, Kao B, Hung E. Mining Frequent Itemsets from Uncertain Data, Springer-Verlag Berlin Heidelberg PAKDD'07. 2007; 4426; 47-58.

[11]. Aggarwal C C, Yan L. Wang Jianyong, Wang Jing., Frequent pattern mining with uncertain data, In Proc. KDD. 2009; 29-37.

[12]. Leung C K S, Carmichael C L. Hao B., Efficient mining of frequent patterns from uncertain data, In Proc. IEEE ICDM Workshops'07. 2007; 489-494.

[13]. Wang K, Tang L, Han J, Liu J. Top down FPGrowth for Association Rule Mining. Proc.Pacific-Asia Conference, PAKDD 2002, 334-340. 2002.

[14]. Abd-Elmegid L A, El-Sharkawi M E, El-Fangary L M, Helmy Y K. Vertical Mining of Frequent Patterns from Uncertain Data, Computer and Information Science. 2010; 3(2); 171–179.

[15]. Tang P, Peterson E A. Mining Probabilistic Frequent Closed Itemsets in Uncertain Databases, 49th ACM Southeast Conference.2011; 86-91.

[16]. Deshpande A, Guestrin C, Madden S R, Hellerstein J M, Hong W. Model-Driven Data Acquisition in Sensor Networks, VLDB; 2004.

[17]. Chen H, Ku W S, Wang H, Sun M T. Leveraging Spatio-Temporal Redundancy for RFID Data Cleansing, In SIGMOD. 2010.

[18]. Pelekis N, Kopanakis I, Kotsifakos E E, Frentzos E, Theodoridis Y. Clustering Uncertain Trajectories, Knowledge and Information Systems. 2010.

[19]. Khare N, Adlakha N, Pardasani K R. Karnaugh Map Model for Mining Association Rules in Large Databases, International Journal of Computer and Network Security. 2009; 1(2); 16–21.

[20]. Lin Y C, Hung C M, Huang Y M, Mining Ensemble Association Rules by Karnaugh Map, World Congress on Computer Science and Information Engineering. 2009; 320–324.

[21]. Zhang Q, Li F, Yi K. Finding frequent items in probabilistic data, In Proc. ACM SIGMOD'08. 2008; 819–832.

[22]. Appell D. The New Uncertainty Principle, Scientific American; 2001.

[23]. Ngai W K, Kao B, Chui C K. Efficient Clustering of Uncertain Data, Proceedings of the Sixth International Conference on Data Mining (ICDM'06), 2006; 2701–2709.

[24]. Leung C K S, Carmichael C L, Hao B.Efficient mining of frequent patterns from uncertain data, In Proc. IEEE ICDM Workshops. 2007; 489-494.

[25]. http://www.stats.gla.ac.uk/steps/glossary/probability.html#probability

[26]. Huang J, Antova L, Koch C, Olteanu D. MayBMS: A probabilistic database management system, in Proc. ACM SIGMOD'09. 2009; 1071–1074.