FREQUENT ITEM SET GENERATION USING AN EFFICIENT MINING ALGORITHM

U.VENKANNA
PG student
Laqshya College of Computer Studies, Kakatiya University, AP, INDIA

Abstract - Frequent patterns recognition is one of the emerging topics in field of data mining. Pattern recognition is one of the major challenges in the field of data mining and knowledge discovery. In our paper we analyzed widely used algorithms to determine frequent patterns and to discover how these algorithms can be used to obtain frequent patterns over the large transactional databases. In our paper we undergone a survey on the frequent itemset mining algorithms such as Apriori algorithm, Frequent Pattern growth algorithm, Rapid Association Rule Mining(RARM),ECLAT algorithm and Associated Sensor Pattern Mining of Data Stream(ASPMS),SMINE.We focused on the strength and weaknesses of these algorithms to obtain frequent patterns among large item sets in database systems.

Keywords - FP Growth algorithm, Rapid Association Rule Mining(RARM),Data mining, Frequent patterns.

I. INTRODUCTION

Frequent pattern mining has been an imperative topic in information mining from numerous years. A striking advance in this field has been made and loads of effective calculations have been intended to look visit designs in a value-based database. Agrawal et al. (1993) right off the bat proposed example mining idea in type of showcase based examination for discovering relationship between items purchased in a market. This idea utilized value-based databases and other information stores so as to separate affiliation's easygoing structures, fascinating connections or visit designs among set of [1]. Frequent patterns are those things, arrangements or substructures that repeat in database transactions with a client indicated recurrence. An itemset with recurrence more prominent than or equivalent to least limit will be considered as a frequent pattern. For instance in market based investigation if the least edge is 30% and bread shows up with eggs and drain multiple occasions or if nothing else multiple times at that point it will be an incessant itemset [2]. Frequent pattern mining can be utilized in an assortment of real world applications. It tends to be utilized in general stores for moving, item situation on racks, for advancement rules and in content looking.We are in an age often referred to as the information age. In this information age, because we believe that information leads to power and success, and thanks to sophisticated technologies such as computers, satellites, etc., we have been collecting tremendous amounts of information. Initially, with the advent of computers and means for mass digital storage, we started collecting and storing all sorts of data, counting on the power of computers to help sort through this amalgam of information. Unfortunately, these massive collections of data stored on disparate structures very rapidly became overwhelming. This initial chaos has led to the creation of structured databases and database management systems (DBMS).The efficient database management systems have been very important assets for management of a large corpus of data and especially for effective and efficient retrieval of particular information from a large collection whenever needed. The proliferation of database management systems has also contributed to recent massive gathering of all sorts of information. Today, we have far more information than we can handle: from business transactions and scientific data, to satellite pictures, text reports and military intelligence. Information retrieval is simply not enough anymore for decision-making. Confronted with huge collections of data, we have now created new needs to help us make better managerial choices. These needs are automatic summarization of data, extraction of the “essence” of information stored, and the discovery of patterns in raw data.

II. FREQUENT ITEMSET MINING ALGORITHMS

2.1 Apriori algorithm

It is given by R. Agrawal and R. Srikant in 1994 for finding frequent item sets in a dataset for Boolean association rule. Name of algorithm is Apriori is because it uses prior knowledge of frequent itemset properties. We apply a iterative approach or level-wise search where k-frequent item sets are used to find k+1 item sets. All nonempty subset of frequent itemset must be frequent. The key concept of Apriori algorithm is its anti-monotonicity of support measure.

2.2 Rapid Association Rule Mining (RARM)

Association rule mining is a well-researched area where many algorithms have been proposed to improve the speed of mining. In this paper, we propose an innovative algorithm called Rapid Association Rule Mining (RARM) to once again break this speed barrier. It uses a versatile tree structure known as the Support-Ordered Trie Itemset (SOTrieIT) structure to hold pre-processed transactional data. This allows RARM to generate large 1-itemsets and 2-itemsets quickly without scanning the database and without candidate 2-itemset generation. It achieves significant speed-ups because the main bottleneck in association rule mining using the Apriori property is the generation of candidate 2-itemsets. RARM has been compared with the classical mining algorithm Apriori and it is found that it outperforms Apriori by up to two orders of magnitude (100 times), much more than what recent mining algorithms are able to achieve.

2.3 FP Growth algorithm

FP growth algorithm is an improvement of apriori algorithm. FP growth algorithm used for finding frequent itemset in a transaction database without candidate generation.

II. FREQUENT PATTERN ANALYSIS

Frequent patterns are itemsets, subsequences, or substructures that appear in a data set with frequency no less than a user-specified threshold. For example, a set of items, such as milk and bread, that appear frequently together in a transaction data set, is a frequent itemset. A subsequence, such as buying first a PC, then a digital camera, and then a memory card, if it occurs frequently in a shopping history database, is a (frequent) sequential pattern. A substructure can refer to different structural forms, such as subgraphs, subtrees, or sublattices, which may be combined with itemsets or subsequences. If a substructure occurs frequently in a graph database, it is called a (frequent) structural pattern. Finding frequent patterns plays an essential role in mining associations, correlations, and many other interesting relationships among data. Moreover, it helps in data indexing, classification, clustering, and other data mining tasks as well. Frequent pattern mining is an important data mining task and a focused theme in data mining research. Abundant literature has been dedicated to this research and tremendous progress has been made, ranging from efficient and scalable algorithms for frequent itemset mining in transaction databases to numerous research frontiers, such as sequential pattern mining, structured pattern mining, correlation mining, associative classification, and frequent pattern-based clustering, as well as their broad applications.

III. ALGORITHM USED

SMine algorithm is an efficient algorithm to find frequent itemsets. The number of database scans is reduced when compared with Apriori algorithm. During the first database scan the number of occurrences of each item is determined and the infrequent ones are discarded. Then the frequent items are counted in each transaction. The transactions are sorted based on the number of frequent items in descending order. Then graph based approach is used to find the frequent item sets.

4.1 Procedure

Step 1: The algorithm scans the database in order to count the number of occurrences of each item to find the candidate 1-itemset with their support count.

Step 2: The set of frequent 1-itemset L1 can then be determined by removing the items having less than the minimum support count. It consists of the candidate 1-itemsets satisfying minimum support. Let the number of frequent 1-itemset be ‘n’.

Step 3: Removes the infrequent items from each transaction and count the number of items in each transaction (item_count). Step 4: The transactions are sorted in descending order based on the item count. Step 5: Create a table called ’M’ with two columns namely ‘no.of.items’ and ‘no.of. transactions’. Let no=n;

Step 6: Adda row with no.of.items = no and no.of.transactions is equal to the number of transactions having item_count >= no. If no >2, then decrement ‘no’ value by 1 and repeat step 6.

Step 7: Select the maximum ‘no.of.items’ from the table M having the ‘no.of.transactions‘ equal to or greater than the minimum support count. Let it be m. Create a directed graph starting from all possible items in m-itemset as the header nodes in the first level, all the possible (m-1) itemset in the second level, all the possible (m-2) itemset in the next level and so on, until 2-itemset.

Step 8 : Get the ‘no.of transactions’ for m-itemset from the table M. Let it be ‘R’. If it is greater than or equal to the minimum support count, then find the support count of each unvisited node of the m-itemset by scanning first ‘R’ transactions. If the set is frequent, mark this node and all its sub nodes as frequent items. If a set is frequent, all its subsets must also be frequent.

Step 9 : Go to the next level .Let m=m-1. Repeat step 8 until m =2. Step 10: All the marked nodes are frequent itemsets.

4.2 Advantages

SMINE algorithm has following advantages

1) Preserves the association information of all itemsets
2) Effective scanning of entire database to obtain frequent patterns to be done to calculate support and confidence measures.
3) Optimizes performance and scalablity
4) Far better than Apriori algorithm
5) Mines more than 300 objects and 10 attributes with an execution time that does not exceed 1200ns.

IV. CONCLUSION

The objective of this study is to review the strengths and weaknesses of the important and recent algorithms in Frequent Pattern Mining (FPM) so that a more efficient FPM algorithm can be developed.After analyzing all the algorithms SMINE proved as a efficient algorithm for FP mining.. In summary, two major problems in FPM have been identified in this research. First, the hidden patterns that exist frequently in a data set become more time consuming to be mined when the amount of data increases. It causes large memory consumption as a result of heavy computation by the mining algorithm. In order to solve these problems, the next stage of the research aims to: (1) formulate an FPM algorithm that efficiently mines the hidden patterns within a shorter run time; (2) formulate the FPM algorithm to consume less memory in mining the hidden patterns; (3) evaluate the proposed FPM algorithm with some existing algorithms in order to ensure that it is able to mine an increased data set within a shorter run time with less memory consumption. By implementing the proposed FPM algorithm, users will be able to reduce the time of decision making, improve the performance and operation, and increase the profit of their organization .

V. REFERENCES

[1] Agarwal RC, Aggarwal CC, Prasad VVV (2001) A tree projection algorithm for generation of frequent item sets. J Parallel Distrib Comput 61(3):350–371

[2]Aggarwal CC (2014) An introduction to Frequent Pattern Mining. In: Aggarwal CC, Han J (eds) Frequent Pattern Mining. Springer, Basel, pp 1–14

[3]Aggarwal CC, Bhuiyan MA, Hasan MA (2014) Frequent Pattern Mining algorithms: a survey. In: Aggarwal CC, Han J (eds) Frequent Pattern Mining. Springer, Basel, pp 19–64

[4]Agrawal R, Srikant R (1994) Fast algorithms for mining association rules. In: Paper presented at the proceedings of the 20th international conference on very large data bases, Santiago

[5]Baralis E, Cerquitelli T, Chiusano S, Grand A (2013) P-Mine: parallel itemset mining on large datasets. In: Paper presented at the 2013 IEEE 29th international conference on data engineering workshops (ICDEW), Brisbane

[6]Chang V (2014) The business intelligence as a service in the cloud. Future Gener Comput Syst 37:512–534CrossRef

[7]Chee C-H, Yeoh W, Tan H-K, Ee M-S (2016) Supporting business intelligence usage: an integrated framework with automatic weighting. J Comput Inf Syst 56(4):301–312

[8]El-Hajj M, Zaiane OR (2003) COFI-Tree mining—a new approach to pattern growth with reduced candidacy generation. In: Paper presented at the workshop on frequent itemset mining implementations (FIMI'03) in conjunction with IEEE-ICDM, Melbourne

[9]Feddaoui I, Felhi F, Akaichi J (2016) EXTRACT: new extraction algorithm of association rules from frequent itemsets. In: Paper presented at the 2016 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM), San Francisco

[10]Gullo F (2015) From patterns in data to knowledge discovery: what data mining can do. Phys Proc 62:18–22CrossRefGoogle Scholar

[11]Gupta B, Garg D (2011) FP-tree based algorithms analysis FPGrowth, COFI-Tree and CT-PRO. Int J Comput Sci Eng 3(7):2691–2699

[12]Gupta A, Tyagi S, Panwar N, Sachdeva S (2017) NoSQL databases: critical analysis and comparison. In: Paper presented at the 2017 international conference on computing and communication technologies for smart nation (IC3TSN), Gurgaon

[13]Han J, Pei J (2000) Mining frequent patterns by pattern-growth: methodology and implications. ACM SIGKDD Explor Newsl: Special issue on "Scalable Data Mining Algorithms", 2(2): 14–20

[14]Han J, Pei J, Yin Y (2000) Mining frequent patterns without candidate generation. ACM SIGMOD Rec 29(2):1–12CrossRefGoogle Scholar

[15]Han J, Kamber M, Pei J (2012) Data mining concepts and techniques. Elsevier, AtlantazbMATH

[16]Hasan MH, Jaafar J, Hassan MF (2014) Monitoring web services' quality of service: a literature review. Artif Intell Rev 42(4):835–850CrossRef

[17]Haupt R, Scholtz B, Calitz A (2015) Using business intelligence to support strategic sustainability information management. In: Paper presented at the 2015 annual research conference on South African institute of computer scientists and information technologists, Stellenbosch

[18]Hoseini MS, Shahraki MN, Neysiani BS (2015) A new algorithm for mining frequent patterns in CanTree. In: Paper presented at the international conference on knowledge-based engineering and innovation, Tehran

[19]Jamsheela O, Raju G (2015) Frequent itemset mining algorithms: a literature survey. In: Paper presented at the 2015 IEEE international advance computing conference (IACC), Banglore

[20]Jesus E, Bernardino J (2014) Open source business intelligence in manufacturing. In: Paper presented at the 18th international database engineering and applications symposium, Porto

[21]King T (2016) Gartner: BI and analytics top priority for CIOs in 2016. Retrieved from https://solutionsreview.com/business-intelligence/gartner-bi-analytics-top-priority-for-cios-in-2016/ Accessed 5 May 2017