Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 37 (2014) 109 – 116

Procedia Computer Science

The 5th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN-2014)

Frequent Pattern Mining Algorithms for Finding Associated Frequent Patterns for Data Streams: A Survey

Shamila Nasreen[a], Muhammad Awais Azam[b], Khurram Shehzad[a], Usman Naeem[c], Mustansar Ali Ghazanfar[a]

[a] *Department of Software Engineering, UET Taxila, 47080, Pakistan*
[b] *Department of Computer Engineering, UET Taxila, 47080, Pakistan*
[c] *School of Architecture, Computing and Engineering, University of East London, United Kingdom*

Abstract

Pattern recognition is seen as a major challenge within the field of data mining and knowledge discovery. For the work in this paper, we have analyzed a range of widely used algorithms for finding frequent patterns with the purpose of discovering how these algorithms can be used to obtain frequent patterns over large transactional databases. This has been presented in the form of a comparative study of the following algorithms: Apriori algorithm, Frequent Pattern (FP) Growth algorithm, Rapid Association Rule Mining (RARM), ECLAT algorithm and Associated Sensor Pattern Mining of Data Stream (ASPMS) frequent pattern mining algorithms. This study also focuses on each of the algorithm's strengths and weaknesses for finding patterns among large item sets in database systems.

© 2014 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/3.0/).

Peer-review under responsibility of the Program Chairs of EUSPN-2014 and ICTH 2014.

Keywords: Frequent Pattern Growth (FP Growth), Rapid Association Rule Mining (RARM), Data Mining, Frequent Patterns

1. Introduction

Frequent pattern mining has been an important subject matter in data mining from many years. A remarkable progress in this field has been made and lots of efficient algorithms have been designed to search frequent patterns in a transactional database. Agrawal et al. (1993) firstly proposed pattern mining concept in form of market based analysis for finding association between items bought in a market. This concept used transactional databases and other data repositories in order to extract association's casual structures, interesting correlations or frequent patterns among set of [1]. Frequent patterns are those items, sequences or substructures that reprise in database transactions with a user specified frequency. An itemset with frequency greater than or equal to minimum threshold will be considered as a frequent pattern. For example in market based analysis if the minimum threshold is 30% and bread appears with eggs and milk more than three times or at least three times then it will be a frequent itemset [2].

Frequent pattern mining can be used in a variety of real world applications. It can be used in super markets for selling, product placement on shelves, for promotion rules and in text searching. It can be used in wireless sensor networks especially in smart homes with sensors attached on Human Body or home usage objects and other applications that require monitoring of user environment carefully that are subject to critical conditions or hazards such as gas leak, fire and explosion [3].These frequent patterns can be used to monitor the activities for dementia patients. It can be seen as an important approach with the ability to monitor activities of daily life in smart environment for tracking functional decline among dementia patients [4].

* Shamila Nasreen. Tel.: +92-05827-961016; fax: +92-05827-961016.
E-mail address: Shamila_nasreen131@yahoo.com

1877-0509 © 2014 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/3.0/).
Peer-review under responsibility of the Program Chairs of EUSPN-2014 and ICTH 2014.
doi:10.1016/j.procs.2014.08.019

In mining pattern stage different techniques are applied to find candidates for frequent patterns and then frequent patterns are generated. There are two main problems with frequent pattern mining techniques. First problem is that the database is scanned many times, second is complex candidate generation process with too many candidate itemset generated. These two problems are efficiency bottleneck in frequent pattern mining. Studies demonstrate that a lot of efforts have been performed for devising best techniques and worth mentioning approaches are Apriori, RARM, ECLAT, FP Growth and ASPMS algorithms.

2. Literature Review

Several algorithms for mining associations have been suggested in the literature work [5] [6] [7] [8] [3] [9] [10] [11] [12] The Apriori algorithm [5] is most widely used algorithm in the history of association rule mining that uses efficient candidate generation process, such that large Itemset generated at k level are used to generate candidates at $k+1$ level. On the other hand, it scans database multiple times as long as large frequent Itemsets are generated. Apriori *TID* generates candidate Itemset before database is scanned with the help of Apriori-gen function. Database is scanned only first time to count support, rather than scanning database it scans candidate Itemset. This variation of Apriori performs well at higher level where as the conventional Apriori performs better at lower levels [6]. Apriori Hybrid is a combination of both the Apriori and Apriori *TID*. It uses apriori *TID* in later passes of database as it outperforms at high levels and Apriori in first few passes of database. DHP (Direct hashing and Pruning) [7] tries to maximize the efficiency by reducing the no of candidates generated but it still requires multiple scans of database. DIC [8] based upon dynamic insertion of candidate items, decrease the number of database scan by dividing the database into intervals of particular sizes. CARMA(Continuous Association Rule Mining Algorithm) proposed in [9] generates more candidate Itemset will less scan of database than Apriori and DIC, however it adds the flexibility to change minimum support threshold. ECLAT [10] with vertical data format uses intersection of transaction ids list for generating candidate Itemset. Each item is stored with its list of Transaction ids instead of mentioning transaction ids with list of items. Sampling algorithm chokes the limitation of I/O overhead by scanning only random samples from the database and not considering whole database. Rapid Association Rule mining (RARM) proposed in [11] generates Large *1- Itemset* and large *2-Itemset* by using a tree Structure called *SOTrieIT* and without scanning database. It also avoids complex candidate generation process for large *1-Itemset* and Large *2-Itemset* that was the main bottleneck in Apriori Algorithm.

Another accomplishment in the development of association rule mining and frequent pattern mining is FP-Growth Algorithm which overcomes the two deficiencies of the Apriori Algorithm [1]. Efficiency of FP-Growth is based on three salient features: (1) A divide-and-conquer approach is used to extract small patterns by decomposing the mining problem into a set of smaller problems in conditional databases, which consequently reduces the search space (2) FP-Growth algorithm avoid the complex Candidate Itemset generation process for a large number of candidate Itemsets, and (3) To avoid expensive and repetitive database scan, database is compressed in a highly summarized, much smaller data structure called FP tree [12]. In [3] a novel tree structure is proposed, called associated sensor pattern stream tree (ASPS-tree) and a new technique, called associated sensor pattern mining of data stream (ASPMS), using sliding window-based associated sensor pattern mining for Wireless Sensor Networks. ASPMS algorithm can extract associated sensor patterns in the current window with frequent pattern (FP)-growth like pattern-growth method after getting useful information from the ASPS-Tree.

3. Comparative Study of Pattern Mining Techniques

Frequent pattern mining techniques have become an obvious need in many real world applications e.g. in market basket analysis, advertisement, medical field, monitoring of patients routines etc. To make a comparison among these algorithms, we use the same transactional database for all algorithms, this transactional database is based on data of a smart home where sensors are installed on daily usage objects and patients while performing their daily tasks, touch these objects and these sensor items are maintained in database. Studies of Frequent pattern mining is acknowledged in the data mining field because of its applicability in mining sequential patterns, structural patterns, mining association rules, constraint based frequent patterns mining, correlations and many other data mining tasks. Efficient algorithms for mining frequent Itemsets are crucial for mining association rules as well as for some other information mining assignments [13]The Problem of mining frequent itemset ascended first as sub-problem of mining association rules [5].

Download English Version:

https://daneshyari.com/en/article/487627

Download Persian Version:

https://daneshyari.com/article/487627

Daneshyari.com