Binary Journal of Data Mining & Networking 4 (2014) 45-48

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

Research Article

ISSN: 2229 -7170

Mining Frequent Item Sets Data Streams using "ÉclatAlgorithm"

Geetanjli Khambra[1] , Pankaj Richhariya[2]

*Corresponding author:

Geetanjli Khambra

[1]Computer Science Department
BITS, Bhopal, India
[2]Computer Science & Engineering
BITS, Bhopal, India

Abstract

Frequent pattern mining is the process of mining data in a set of items or some patterns from a large database. The resulted frequent set data supports the minimum support threshold. A frequent pattern is a pattern that occurs frequently in a dataset. Association rule mining is defined as to find out association rules that satisfy the predefined minimum support and confidence from a given data base. If an item set is said to be frequent, that item set supports the minimum support and confidence. A Frequent item set should appear in all the transaction of that data base. Discovering frequent item sets play a very important role in mining association rules, sequence rules, web log mining and many other interesting patterns among complex data. Data stream is a real time continuous, ordered sequence of items. It is an uninterrupted flow of a long sequence of data. Some real time examples of data stream data are sensor network data, telecommunication data, transactional data and scientific surveillances systems. These data produced trillions of updates every day. So it is very difficult to store the entire data. In that time some mining process is required. Data mining is the non-trivial process of identifying valid, original, potentially useful and ultimately understandable patterns in data. It is an extraction of the hidden predictive information from large data base. There are lots of algorithms used to find out the frequent item set. In that Apriori algorithm is the very first classical algorithm used to find the frequent item set. Apart from Apriori, lots of algorithms generated but they are similar to Apriori. They are based on prune and candidate generation. It takes more memory and time to find out the frequent item set. In this paper, we have studied about how the éclat algorithm is used in data streams to find out the frequent item sets. Éclat algorithm need not required candidate generation.

Keywords: Association rules mining, Data mining, Data streams, Éclat algorithm, frequent pattern mining.

Introduction

Data stream in mining is the process of extracting knowledge from, continuous rapid growth of data. Data stream is a prearranged series of items that arrives in timely order [9]. It is impossible to store the data in which item arrives. To apply data mining algorithm directly to streams instead of storing them before in a database. In data streams the items are represented by record structure i.e. each individual data items may be relational tuples. Call records, web page visits, sensor reading are some examples of tuples in data streams. The hasty growth of uninterrupted data has many challenges to store, computation and communication capabilities in computing system.

In data stream data enters at high speed and continuous way. It is not possible to store them in a data warehouse. To identify a nugget that is some chunk of information in the database and extracting this information in some meaningful way is known as data mining [1] [3]. In that time some techniques are required to process the large data base. So in data streams, data mining techniques help to find interesting patterns and anomalies in the data. Data mining techniques plays a vital role in many large organizations. But nowadays, many new techniques and algorithms are used for data streams without dropping the events. Data stream algorithms are designed with clear focus on the evolution of the underlying data.

This paper will focus on the following sections. In Section 2, we present an overview of association rule. Section 3 discusses the various models of data streams. Section 4 gives the overview of frequent pattern mining. Section 5 discusses the éclat algorithm and its analysis in frequent pattern mining. Experimental results are discussed in section 6. The conclusion and future work of this paper is converses in Section 7.

The main objective of the association rule is to discover all the rules that have the support and confidence greater than or equal to minimum support and confidence. While using this association rule, the user can skip the lower amount of data in the huge data base. The association rule is used to help the retailer to improve their marketing strategies, to recognize "which items are frequently purchased by clients". It also helps in inventory management, sales promotion strategies etc.

Association Rule

An association rule is an expression which has the form of X->Y, where X and Y are the subsets of A [1] [4]. An association has two measures namely called as support and confidence. Support which means how often items occur together as a percentage of total transaction. Confidence which is used to measure how much a particular item is dependent on another. For example, an association rules states that T consists of 500000 transactions; 20000 transactions of these contains chips; 30000 transactions contain beer. 10000 transactions contain both chips and beer.

In this transaction the support is 3% i.e. 10000/500000. The confidence is 33.33% calculated by using the calculation 10000/30000. An item set in association rule is called frequent if its support (frequency) beats a given minimal support (frequency) threshold.

Association rule mining process is a two step process. They are, [1]:

First is to find all frequent item sets in a data

By definition, each of these item sets will occur at least as frequently as it satisfies the minimum support count, minimum confidence process.

Second the user will generate strong association rules between the frequently occurred item sets.

According to the definition this rule must satisfy same minimum support and confidence threshold.

The main problem under this association rule is, if the frequent item set is large, the combination of the item sets will be very huge. So it can create many problems. There are two conditions are used for solving this problem. They are [1]:

Maximal frequent item set: an item set is maximal,

if none of its immediate supersets is frequent.

Closed frequent item sets: an item set is closed, if none of its immediate supersets has the same support as the item sets.

Data Stream

Mining data stream is the process of finding hidden knowledge from continuous, rapid data records. Data arrives faster, so it is a very difficult task to mine that data. Stream mining algorithms typically need to be designed so that the algorithm works with one pass of the data. Mining data streams are a computational challenge to perform tasks because of the additional algorithmic restraints created by the large volume of data. In addition, the problem of temporal locality leads to a number of unique mining challenges in the data stream case.

Some solutions have been proposed for data stream mining problems. Summarization techniques are used to produce approximation solution from large database. Summarization techniques refer the process of transforming data to a suitable form for stream data analysis. This can be done by summarizing the entire stream or some subset of incoming stream.

These solutions can be categorized into two types

[9] [12],

Data-based stream model

Task-based stream model

A. Data-based stream model

Data-based techniques refer to summarize the whole dataset or choosing a subset of the incoming stream to be analyzed.

B. Task-based stream model

Task-based techniques are opposite to data based model because these types modify or change the existing algorithm or create new ones to address the challenges in data stream

Frequent Item Set Mining Methods

An item set is a collection of one or more items in a transaction. Example T= {chips, beer, candy} is an item set.

An item set's threshold value supports the minimum support and confidence is known as frequent item set.

Finding frequent item set is not an easy task. Patterns that are frequently in data are known as frequent item set. There are many kinds frequent patterns are in use. They are [2] [12],

Frequent item set -> ex: set of items like milk, bread appears together in a transactional data is known as frequent item set. (Items frequently appears together)

Subsequence item set -> ex: PC, digital camera, memory card which means a person first buys a PC, then a digital camera and then a memory card. If these items are appearing together in a shopping database, it is a sequential pattern.

Sub structure items -> ex: graph, tree, and lattice. Different structural forms combined with item sets or subsequence is called as substructure.

In data streams data arrives continuously and the volume of transaction will be very huge. There are lots of techniques used in data streams. A user can uses different types of window model over data streams to find the frequent item sets.

Land mark window model

Damped window model

Sliding window model

Land Mark Window Model

This model uses some time point of the previous transaction to the current time point transaction. This window model used to find the frequent item sets over the entire data streams.

Damped Window Model

Damped window model gives some weight to the recently arrived transaction. It automatically omits the previous transaction, when a new transaction arrives. This window model is mainly used in nowadays.

Sliding Window Model

The time changes this window will change the size and moves along with the current time point. It sets some threshold value to each window. If any new item arrive this window will automatically slide once and find the frequent item sets.

Eclat Algorithm In Data Streams -Analysis

Equivalence Class Clustering and bottom up Lattice Traversal is known as ECLAT algorithm. This algorithm is also used to perform item set mining. It uses tid set intersection that is transaction id intersection to compute the support of a candidate item set for avoiding the generation of subsets that does not exist in the prefix tree. For each item store a list of transaction id. It uses vertical data layout.

For calculating the 1 candidate frequent item sets, this algorithm assumes that the vertical t-id list data base is given and for each item it simply reads it corresponding tid list from the given data base and incrementing the items support for each entry. Éclat algorithm attempts to improve the fastness of the support computations. Compared with other algorithms like Apriori, FP-growth etc, it does not create the candidate item sets. This algorithm scans the data base only once and creates the vertical data base, which identifies each item in the list of transactions that supports the items.

Consider the following example. The table 1 gives the transaction in a particular time period. There are five item sets and nine transactions in that data base.

Table 1: transaction data base

Windows	Transaction
W1	1 0 0 0 1 0 1
W2	1 0 0 1 0 0 1 0
W3	1 1 0 0 1 0 0 0
W4	0 1 0 1 1 0 1 0
W5	0 0 1 0 1 1 1

With the help of the éclat algorithm we can put tid to each transaction in table 1. This algorithm checks the Transaction data in item wise. It puts id to each transaction.

Conclusion And Future Work

In this paper we have studied the concept of éclat algorithm and its experimental results. In the course of this algorithm we may discover, when the window size increases the execution time also increases. Moreover, this algorithm demonstrates that the runtime difference decreased when the minimum threshold value increased.

References

[1] Agrawal R, Imielinski T, and Swami AN. "Mining association rules between sets of items in large databases," in *ACM SIGMOD International Conference on Management of Data*, Washington, 1993.

[2] Agrawal R, and Srikant R. "Fast algorithms for mining association rules," in *20th International Conference on Very Large Data Bases*, Washington, 1994.

[3] Han J, Pei J, and Yin Y. "Mining frequent patterns without candidate generation," in *ACM SIGMOD International Conference on Management of Data*, Texas, 2000.

[4] Zaki MJ, Parthasarathy S, Ogihara M, and Li W. "New algorithms for fast discovery of association rules," in *Third International Conference on Knowledge Discovery and Data Mining*, 1997.

[5] KG MJ. Zaki, "Fast vertical mining using diffsets," in The nineth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2003.

[6] Paul W. Purdom, Dirk Van Gucht, and Dennis P. Groth, "Average case performance of the apriori algorithm," vol. 33, p. 1223–1260, 2004.

[7] Orlando S, Palmerini P, Perego R, and Silvestri F. "Adaptive and resource-aware mining of frequent sets," in *Proceedings of the 2002 IEEE International Conference on Data Mining*, 2002.

[8] Shenoy P, Haritsa JR, Sudarshan S, Bhalotia G, Bawa M, and Shah D. "Turbo-charging vertical mining of large databases," in *ACM SIGMOD International Conference on Management of Data*, 2000.

[9] Zheng Z, Kohavi R, and Mason L. Real World Performanceof Association Rule Algorithms. In: Proc.7th Int. Conf. on Knowledge Discovery and Data Mining(SIGKDD'01). ACM Press, New York, NY, USA2001.

[10] "Data Streams: An Overview and Scientific Applications" Charu C. Aggarwal .

[11] Han J, Kamber M. "Data Mining Concepts and Techniques", Morgan Kaufmann Publishers, 2006 .

[12] Agarwal R, Srikant R. " fast algorithms for mining Association rules" 20th Int Conference.

[13]. Pauray SM. Tsai, "Mining frequent item sets in data streams using the weighted sliding window model", Elsevier publication 2009.

[14]. Syed Khairuzzaman Tabeer, Chowdary Farha ahmed, Byeong-Soo Jeong, Young Koo Lee"Efficient frequent pattern mining over datastreams" 2008.

[15]. Tanbeer S K, Ahmed CF, Jeong B-S, and Lee Y-K. 2008. "CP-tree: a tree structure for single-pass frequent pattern mining"S. In Proc.of PAKDD, Lect Notes Artif Int, 1022-1027.

[16]. Yo unghee Kim, Won Young Kim and Ungmo Kim "Mining frequent item sets with normalized weight in continuous data streams". Journal ofinformation processing systems. 2010