

Discovering Association Rules in Data Streams Based On Closed Pattern Mining

Nan Jiang
School of Computer Science
The University of Oklahoma
Norman, OK 73019, USA
1(405)325-9821
nan_jiang@ou.edu

ABSTRACT

As more and more applications such as traffic modeling, military sensing and tracking, online data processing generate a large amount of data streams every day, efficient knowledge discovery of data streams is an emerging active research area in data mining with broad applications. Due to the unique features of data stream, traditional data mining techniques which require multiple scans of the entire data sets can not be applied directly to mine stream data which usually allows only one scan and demands fast response time. This research proposes methods to first mine closed patterns in data streams, and then generate association rules based on closed patterns which contain non-redundant and complete information and are more useful for data analysis. The proposed mining technique in data streams is then be applied to a sensor network database to identify relationships between sensor readings for missing data estimation purpose.

Categories and Subject Descriptors

H.2.8 [Database Management]: Database Applications – Data Mining.

General Terms

Algorithms, Performance, Experimentation.

Keywords

Data Streams, Closed Frequent Itemsets, Association Rules

1. MOTIVATIONS

As the number of applications on mining data streams grows rapidly, there is an increasing need to perform association rule mining on stream data. One example application of data stream association rule mining is to estimate missing data in sensor networks [6]. Another example is to predict frequency estimation of Internet packet streams [5]. In the MAIDS project [4], this technique is used to find alarming incidents from data streams. Association rule mining can also be applied to monitor manufacturing flows to predict failure or generate reports based on web log streams, and so on. A data stream is an ordered sequence of transactions that arrives in

timely order. Different from data in traditional static databases, data streams have the following characteristics. First, they are continuous, unbounded, and usually come with high speed. Second, the volume of data streams is large and usually with open end. Third, the data distribution in streams is usually changing with time.

Traditional association rule mining algorithms are developed to work on static data and, thus, can not be applied directly to mine association rules in stream data. The first recognized frequent itemsets mining algorithm for traditional databases is Apriori [1]. After that, many other algorithms based on the ideas of Apriori were developed for performance improvement [2, 7]. Apriori-based algorithms require multiple scans of the original database, which leads to high CPU and I/O costs. Therefore, they are not suitable for a data stream environment, in which data can be scanned only once. Another category of association rule mining algorithms for traditional databases proposed by Han and Pei [8] are those using a frequent pattern tree (FP-tree) data structure and an FP-growth algorithm which allows mining of frequent itemsets without generating candidate itemsets. Compared with Apriori-based algorithms, it achieves higher performance by avoiding iterative candidate generations. However, it still can not be used to mine association rules in data streams since the construction of FP-tree requires two scans of data.

In view of those challenges, we aim to develop an efficient association mining algorithm in data streams based on closed pattern mining as they provide complete and condensed pattern information. Also, according to [1], any subset of a frequent itemset is also frequent. Thus, algorithms that mine all frequent itemsets often suffer from the problem of combinatorial explosion. The combinatorial explosion problem of mining frequent itemsets becomes even more serious in the streaming environment because of the huge amount of data. As a result, we cannot afford to keep track of all itemsets or even all frequent itemsets due to the time and space constraints. However, any omission may prevent us from discovering future frequent itemsets. Thus, the challenge lies in designing a compact data structure, which can keep track of all frequent itemsets over data streams. We achieve this by storing only the closed itemsets, the small subset of all frequent itemsets in memory. Furthermore, association rules generated from closed frequent itemsets contain non-redundant and complete information [11] which are more useful for data analysis. This technique is then applied to discover the relationships between sensors in a sensor network database and, then, data missing by a sensor

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Proceedings of SIGMOD2007 Ph.D. Workshop on Innovative Database Research 2007(IDAR2007), June 10, 2007, Beijing, China

is estimated using the data generated by its related sensors in the sensor network database.

2. PROPOSED TECHNIQUES

Here we briefly describe our proposed CFI-Stream algorithm, an efficient algorithm for mining closed itemsets in data streams and its in-memory data structure, called DIrect Update (DIU) tree to perform the closure checking online over a data stream sliding window. Please refer to the detail discussions of the algorithm in our paper [9].

When a transaction arrives or leaves the current data stream sliding window, the algorithm checks each itemset in the transaction on the fly and updates the associated closed itemsets' supports. Current closed itemsets are maintained and updated in real time in the DIU tree. The closed frequent itemsets can be output at any time at users' specified thresholds by browsing the DIU tree.

We use a lexicographical ordered direct update tree (DIU) to maintain the current closed itemsets. Each node in the DIU tree represents a closed itemset. There are k levels in the DIU tree, each level i stores the closed i -itemsets. The parameter k is maximum length of the current closed itemsets. Each node in the DIU tree stores a closed itemset, its current support information, and the links to its immediate parent and children nodes. Figure 1 illustrates the DIU tree after the first four transactions arrive. The support of each node is labeled in the upper right corner of the node itself. The figure shows that currently there are four closed itemsets C, AB, CD, and ABC in the DIU tree, and their associated supports are 3, 3, 1, and 2.

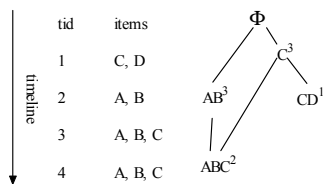


Figure 1. The lexicographical ordered direct update tree

Based on the discovered frequent closed itemsets, we can generate a basic set of rules, from which all other association rules can be inferred. Thus only a small and understandable set of rules need to be presented to the user, which can later selectively derive other rules of interest. We define the non-redundant association rules as follows: If $X_1 \xrightarrow{s,c} X_2$, $X_3 \xrightarrow{s,c} X_4$, $X_1 \subseteq X_3$, and $X_4 \subseteq X_2$, we say that $X_3 \xrightarrow{s,c} X_4$ is redundant [3].

We prove that only immediate connected closed itemsets in the DIU tree need to be considered to generate association rules. Rest association rules can be derived. We define the most minimal itemset X_1 is the itemset that satisfies with $C(X_1) = X_2$, which is called one of X_2 's generators and show that among rules that are equivalent to $C(X_1) \xrightarrow{s,c} C(X_2)$, only rules with the form of $X \xrightarrow{s,c} C(X)$ and X is

$C(X)$'s minimum generator are non-redundant. Rest rules are redundant.

The association rule mining based on the derived closed itemsets forms the foundation for our data estimation algorithm, CARM [10]. Instead of generating all possible association rules, we generate the rules that have strong relationships with the current round of sensor readings where one or more readings are missing. We achieve this through browsing the DIU tree, which stores all of the closed itemsets. Based on the users' specified support and confidence thresholds, we find out rules through paths (links) of closed itemsets that suit the users' needs, i.e., satisfy the users' specified support and confidence thresholds. The mining process is online and incremental, which is especially beneficial when users have different specified thresholds in their online queries.

Our performance studies show that our proposed mining algorithm is able to mine data streams online with both time and space efficiency independent of support information, and it can adapt to the concept-drifting in data streams. The CARM algorithm offers an online method to derive association rules based on the discovered closed itemsets, and imputes the missing values based on the derived association rules. It can find out the relationships between multiple sensors not only when they report the same sensor readings but also when they report different sensor readings, greatly improves the estimation accuracy and reduces the number of cases that cannot be estimated.

3. REFERENCES

- [1] Rakesh Agrawal, Tomasz Imielinski, Arun Swami; Mining Association Rules between Sets of Items in Massive Databases; Int'l Conf. on Management of Data; May 1993.
- [2] Rakesh Agrawal, Ramakrishnan Srikant; Fast Algorithms for Mining Association Rules; Int'l Conf. on Very Large Databases; September 1994.
- [3] Yves Bastide, Nicolas Pasquier, Rafik Taouil, Gerd Stumme, Lotfi Lakhal; Mining Minimal Non-redundant Association Rules Using Frequent Closed Itemsets; First International Conference on Computational Logic; Pages: 972 - 986; July 2000.
- [4] Y. Dora Cai, Greg Pape, Jiawei Han, Michael Welge, Loretta Auvil; MAIDS: Mining Alarming Incidents from Data Streams; Int'l Conf. on Management of Data; June 2004.
- [5] Erik D. Demaine, Alejandro Lopez-Ortiz, J. Ian Munro; Frequency Estimation of Internet Packet Streams with Limited Space; European Symposium on Algorithms; September 2002.
- [6] Mihail Halatchev and Le Gruenwald; Estimating Missing Values in Related Sensor Data Streams; Int'l Conf. on Management of Data; January 2005.
- [7] Jiawei Han, Guozhu Dong, Yiwen Yin; Efficient mining of partial periodic patterns in time series database; IEEE Int'l Conf. on Data Mining; March 1999.
- [8] Jiawei Han, Jian Pei, Yiwen Yin; Mining Frequent Patterns without Candidate Generation; Int'l Conf. on Management of Data; May 2000.
- [9] N. Jiang and L. Gruenwald, "CFI-Stream: Mining Closed Frequent Itemsets in Data Streams", ACM SIGKDD intl. conf. on knowledge discovery and data mining, 2006.
- [10] N. Jiang and L. Gruenwald, "Estimating Missing Data in Data Streams," accepted by the 12th International Conference on Database Systems for Advanced Applications, April 2007.
- [11] Mohammed J. Zaki; Generating non-redundant association rules; ACM SIGKDD international conference on Knowledge discovery and data mining; 2000.