Chapter 14
An Approximate Approach for Maintaining Recent Occurrences of Itemsets in a Sliding Window over Data Streams

Jia-Ling Koh
National Taiwan Normal University, Taiwan

Shu-Ning Shin
National Taiwan Normal University, Taiwan

Yuan-Bin Don
National Taiwan Normal University, Taiwan

ABSTRACT
Recently, the data stream, which is an unbounded sequence of data elements generated at a rapid rate, provides a dynamic environment for collecting data sources. It is likely that the embedded knowledge in a data stream will change quickly as time goes by. Therefore, catching the recent trend of data is an important issue when mining frequent itemsets over data streams. Although the sliding window model proposed a good solution for this problem, the appearing information of patterns within a sliding window has to be maintained completely in the traditional approach. For estimating the approximate supports of patterns within a sliding window, the frequency changing point (FCP) method is proposed for monitoring the recent occurrences of itemsets over a data stream. In addition to a basic design proposed under the assumption that exact one transaction arrives at each time point, the FCP method is extended for maintaining recent patterns over a data stream where a block of various numbers of transactions (including zero or more transactions) is inputted within a fixed time unit. Accordingly, the recently frequent itemsets or representative patterns are discovered from the maintained structure approximately. Experimental studies demonstrate that the proposed algorithms achieve high true positive rates and guarantees no false dismissal to the results yielded. A theoretic analysis is provided for the guarantee. In addition, the authors' approach outperforms the previously proposed method in terms of reducing the run-time memory usage significantly.

DOI: 10.4018/978-1-60566-748-5.ch014
1. INTRODUCTION

The strategies for mining frequent itemsets in static databases have been widely studied over the last decade such as the Apriori (Agrawal & Srikant, 1994), DHP (Park, Chen, & Yu, 1995), and FP-growth (Han et al., 2004) algorithms. Recently, the data stream, which is an unbounded sequence of data elements generated at a rapid rate, provides a dynamic environment for collecting data sources. It is considered that the main restrictions of mining data streams include scanning data in one pass and performing the mining process within a limited memory usage.

Since it is not feasible to store the past data in data streams completely, a method for providing approximate answers with accuracy guarantees is required. The hash-based approach was proposed in (Jin et al., 2003), in which each item in a data stream owns a respective list of counters in a hash table, and each counter may be shared by more than one item. A new novel algorithm, called hCount, was provided to maintain frequent items over a data stream and support both insertion and deletion of items with a less memory space. Lossy-counting is the representative approach for mining frequent itemsets from data streams (Manku & Motwani, 2002). Given an error tolerance parameter $\varepsilon$, the Lossy-counting algorithm prunes the patterns with supports being less than $\varepsilon$ from a pool of monitored patterns such that the required memory usage is reduced. Consequently, the frequency of a pattern is estimated by compensating the maximum number of times that the pattern could have occurred before being inserted into the the pool of monitored patterns. It is proved that no false dismissal occurs with Lossy-counting algorithm. Moreover, for each pattern, the error rate of its estimated frequency is guaranteed not to exceed a given error tolerance parameter.

Although the restriction of memory usage was considered in the two works introduced previously, time sensitivity is another important issue when mining frequent itemsets from data streams. It is likely that the embedded knowledge in a data stream will change quickly as time goes by. In order to catch the recent trend of data, the estDec algorithm (Chang & Lee, 2003) decayed the old occurrences of each itemset as time goes by to diminish the effect of old transactions on the mining result of frequent itemsets over a data stream. The above approach provided time-sensitive mining for long-term data. However, in certain applications, it is interested only the frequent patterns discovered from the recently arriving data within a fixed time period. Under the assumption that exact one transaction arrives at each time unit, the sliding window method (Chang & Lee, 2004) defined the current sliding window to consist of the most recently coming $w$ transactions in a data stream according to a given window size $w$. Consequently, the recently frequent itemsets were defined to be the frequent itemsets mined from the current sliding window. In addition to maintain the occurrence for the new transaction, the oldest transaction has to be removed from the maintained data structure when the window is sliding. However, all the transactions in the current sliding window need to be maintained in order to remove their effects on the current mining result when they are beyond the scope in the window.

In (Lin et al., 2005), a time-sensitive sliding window approach was also proposed for mining recently frequent itemsets within the current sliding window in a data stream. However, a general assumption that a block of various numbers of transactions (zero or more transactions) is inputted into the data stream at each time unit was adopted. Accordingly, the recently frequent itemsets were discovered from the most recent $w$ blocks of transactions. For each block of transactions, the frequent itemsets in the block were found and all possible frequent itemsets in the sliding window were collected in a PFP (Potential Frequent-itemset Pool) table. For each newly inserted pattern, the maximum number of possible lost counts was estimated. Moreover, a discount-