Available online at www.sciencedirect.com
SCIENCE DIRECT®
Information Sciences 176 (2006) 1986–2015
INFORMATION SCIENCES
AN INTERNATIONAL JOURNAL
www.elsevier.com/locate/ins

A false negative approach to mining frequent itemsets from high speed transactional data streams

Jeffrey Xu Yu [a,*], Zhihong Chong [b], Hongjun Lu [c], Zhenjie Zhang [d], Aoying Zhou [b]

[a] *Department of Systems Engineering and Engineering Management, Chinese University of Hong Kong, Shatin, New Territories, Hong Kong, China*
[b] *Fudan University, Shanghai, China*
[c] *Hong Kong University of Science and Technology, Hong Kong, China*
[d] *National University of Singapore, Singapore, Singapore*

Abstract

Mining frequent itemsets from transactional data streams is challenging due to the nature of the exponential explosion of itemsets and the limit memory space required for mining frequent itemsets. Given a domain of I unique items, the possible number of itemsets can be up to $2^I - 1$. When the length of data streams approaches to a very large number N, the possibility of an itemset to be frequent becomes larger and difficult to track with limited memory. The existing studies on finding frequent items from high speed data streams are false-positive oriented. That is, they control memory consumption in the counting processes by an error parameter ϵ, and allow items with support below the specified minimum support s but above $s - \epsilon$ counted as frequent ones. However, such false-positive oriented approaches cannot be effectively applied to frequent itemsets mining for two reasons. First, false-positive items found increase the number

* Corresponding author. Tel.: +852 2609 8309; fax: +852 2603 5505.
E-mail addresses: yu@se.cuhk.edu.hk (J.X. Yu), zhchong@fudan.edu.cn (Z. Chong), luhj@cs.ust.hk (H. Lu), zhangzh2@comp.nus.edu.sg (Z. Zhang), ayzhou@fudan.edu.cn (A. Zhou).

0020-0255/$ - see front matter
doi:10.1016/j.ins.2005.11.003

of false-positive frequent itemsets exponentially. Second, minimization of the number of false-positive items found, by using a small ϵ, will make memory consumption large. Therefore, such approaches may make the problem computationally intractable with bounded memory consumption. In this paper, we developed algorithms that can effectively mine frequent item(set)s from high speed transactional data streams with a bound of memory consumption. Our algorithms are based on Chernoff bound in which we use a running error parameter to prune item(set)s and use a reliability parameter to control memory. While our algorithms are false-negative oriented, that is, certain frequent itemsets may not appear in the results, the number of false-negative itemsets can be controlled by a predefined parameter so that desired recall rate of frequent itemsets can be guaranteed. Our extensive experimental studies show that the proposed algorithms have high accuracy, require less memory, and consume less CPU time. They significantly outperform the existing false-positive algorithms.

Keywords: Data stream; Frequent pattern mining; Memory minimization

1. Introduction

Recently, data streams emerged as a new data type that attracted great attention from both researchers and practitioners. A data stream is essentially a virtually unbounded sequence of data items arriving at a rapid rate. Since data items arrive continuously, it is only feasible to store certain form of synopsis (in memory or disk) rather than the raw data for analysis or information extraction. It is also infeasible to multiple scan the original data to build such synopsis because of the massive volume as well as the rapid arrival rate. Research work related to data streams boils down to the problem of finding the right form of synopsis and related construction algorithms so that the required statistics or patterns can be obtained with a bounded error for unbounded input data items with limited memory. A large amount of work has been reported for various statistics and patterns, including simple aggregates and statistics such as maximum, minimum, average, median values and quantiles as well as complex patterns such as decision trees, clusters, and frequent itemsets.

In this paper we study the problem of mining frequent item(set)s (or patterns) from high speed *transactional data streams*. Manku and Motwani gave an excellent review of wide range applications for the problem of frequent data stream pattern mining [12]. The problem can be stated as follows. Let $I = \{x_1, x_2, \ldots, x_n\}$ be a set of items. An itemset is a subset of items I. A transactional data stream, $\mathscr{D}$, is a sequence of incoming transactions, $(t_1, t_2, \ldots, t_N)$, where a transaction t_i is an itemset and N is a unknown large number of transactions that will arrive. The number of transactions in $\mathscr{D}$ that contain X is