Approximate Data Mining for Sliding Window Based Data Streams

Kuo-Cheng Yin      Yu-Lung Hsieh     Don-Lin Yang

Department of Information Engineering and Computer Science, Feng Chia University
Taichung 407, Taiwan, ROC
{inn0206, yuhlong.hsieh}@gmail.com, dlyang@fcu.edu.tw

Received 7 February 2012; Revised 13 June 2012; Accepted 22 June 2012

Abstract. In the sliding window model of continuous dynamic data streams, the real-time process and update is an important issue for association rule mining. The existing researches deal with the problem by using specific data structures to retain the scanned data. However, if the next window slot contains any new frequent items, all the data must be rescanned to generate itemsets containing the new frequents. It is prohibitive to read the data twice for time-critical mining of continuous data streams. In order to meet the requirement of scanning data only one time, we propose a new approximate data stream mining algorithm (ADSMiner) using an extended FP-tree (EFP-tree) to save the current frequent-patterns. The EFP-tree not only records the frequent itemsets, but also keeps the counts of each itemset in the panes. If any new 1-itemset becomes frequent after the old data is replaced by the new data, there is no need to re-read the data. Instead, it is just added to the EFP-tree. When the order of the frequent 1-itemsets sequence changes, we use the Longest Common Subsequence method to locate the nodes requiring adjustment and maintain the structure of EFP-tree efficiently. The results of experiment show that our approach performs well as we expected on various datasets.

Keywords: association rule, FP-tree, data stream, sliding window, approximate mining

Acknowledgement

This research was supported by the National Science Council, Taiwan, under grant NSC 100-2221-E-035-103.

References


[6] J. Han, J. Pei, Y. Yin, "Mining Frequent Patterns without Candidate Generation," in Proceedings of SIGMOD, pp. 1-12, 2000.


[13] J. Han, J. Pei, Y. Yin, R. Mao, "Mining Frequent Patterns without Candidate Generation: A Frequent-pattern Tree Approach," *Data Mining and Knowledge Discovery*, Vol. 8, No. 1, pp. 53-87, Jan. 2004.


