# 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

[1] P.S.M. Tsai, "Mining Frequent Itemsets in Data Streams Using the Weighted Sliding Window Model," *Expert Systems with Applications*, Vol. 36, No. 9, pp. 11617-11625, 2009.

[2] C.H. Lin, D.Y. Chiu, Y.H. Wu, A.L.P. Chen, "Mining Frequent Itemsets from Data Streams with a Time-sensitive Sliding Window," in *Proceedings of SIAM International Data Mining Conference*, pp. 68-79, 2005.

[3] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules," in *Proceedings of the 20th International Conference on Very Large Data Bases*, pp. 487-499, 1994.

[4] Y.J. Tsay and J.Y. Chiang, "CBAR: An Efficient Method for Mining Association Rules," *Knowledge-Based Systems*, Vol. 18, No. 4, pp. 99-105, 2005.

[5] W. Song, B. Yang, Z. Xu, "Index-BitTableFi: An Improved Algorithm for Mining Frequent Itemsets," *Knowledge-Based Systems*, Vol. 21, No. 6, pp. 507-513, 2008.

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

[7] M. Song and S. Rajasekaran, "A Transaction Mapping Algorithm for Frequent Itemsets Mining," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 18, No. 4, pp. 472-481, 2006.

[8] P.S.M. Tsai and W. Zhuang, "Data Mining for Library Borrowing History Records Based on the Weighted Sliding Window Model," *Journal of Computers*, Vol. 17, No. 4, pp. 79-96, 2006.

[9] B. R. Dai and M.S. Chen, "Constrained Clustering for the Evolving Data Stream," *Journal of Computers*, Vol. 17, No. 4, pp. 37-51, 2007.

[10] J. Cheng, Y. Ke, W. Ng, "A Survey on Algorithms for Mining Frequent Itemsets over Data Streams," *Knowledge and Information Systems*, Vol. 16, pp. 1-27, 2007.

[11] H.F. Li and S.Y. Lee, "Mining Frequent Itemsets over Data Streams Using Efficient Window Sliding Techniques," *Expert Systems with Applications*, Vol. 36, No. 2, pp. 1466-1477, 2009.

[12] S.K. Tanbeer, C.F. Ahmed, B.S. Jeong, Y.K. Lee, "Sliding Window-based Frequent Pattern Mining over Data Streams," *Information Sciences*, Vol. 179, pp. 3843-3865, Nov. 2009.

[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.

[14] J.L. Koh and S.F. Shieh, "An Efficient Approach for Maintaining Association Rules Based on Adjusting FP-tree Structures," *Proceedings of DASFAA*, pp. 417-424, 2004.

[15] C. Giannella, J. Han, J. Pei, X. Yan, P.S. Yu, "Mining Frequent Patterns in Data Streams at Multiple Time Granularities," in *Proceedings of the NSF Workshop on Next Generation Data Mining*, pp. 191-212, 2002.

[16] T.P. Hong, C.W. Lin, Y.L. Wu, "Incrementally Fast Updated Frequent Pattern Trees," *Expert Systems with Applications*, Vol. 34, No. 4, pp. 2424-2435, 2008.

[17] S.K. Tanbeer, C.F. Ahmed, B.S. Jeong, Y.K. Lee, "CP-tree: A Tree Structure for Single-pass Frequent Pattern Mining," in *Proceedings of PAKDD 2008*, pp. 1022-1027, 2008.

[18] L. Bergroth, H. Hakonen, T. Raita, "A Survey of Longest Common Subsequence Algorithms," *SPIRE, A Coruna, Spain*, pp. 39-48, 2000.

[19] R. Kohavi, C. Brodley, B. Frasca, L. Mason, Z. Zheng, "KDD-cup 2000 Organizers’ Report: Peeling the Onion," *SIGKDD Explorations*, Vol. 2, No. 2, pp. 86-98, 2000. <http://www.ecn.purdue.edu/KDDCUP>.