Mining Frequent Items in a Stream Using Flexible Windows

Toon Calders (1), Nele Dexters (2) and Bart Goethals (2)
(1) Eindhoven Technical University, The Netherlands
(2) ADReM Research Group, University of Antwerp, Belgium

We study the problem of finding frequent items in a continuous stream of items. Mining frequent items over a stream of items presents interesting new challenges over traditional mining in static databases. The stream can only be scanned once, and hence if an item is passed, it can not be revisited, unless it is stored in main memory. Storing large parts of the stream, however, is not possible because the amount of data passing by is typically huge.

Most previous work on mining frequently occurring items from a stream either focusses on (1) the whole stream, (2) on only the most recent items in a window of fixed length [1, 2, 4], or (3) where a time-decaying factor fades out the history of the stream [3]. In many applications, however, it is not possible to fix a window length or a decay factor that is most appropriate for every item at every timepoint in an evolving stream. For example, consider a large retail chain of which sales can be considered as a stream. Then, in order to perform market basket analysis, it is very difficult to choose in which period of the collected data you are particularly interested. For many products, the amount sold depends highly on the period of the year. In summer time, e.g., sales of ice cream increase. During the world cup, sales of beer increase. Such seasonal behavior of a specific item can only be discovered when choosing the correct window size for that item, but this size can then also hide a similar behavior of other items. Therefore, a new frequency measure is introduced, based on a flexible window length. We propose to consider for each item the window in which it has the highest frequency. More specifically, we define the current frequency of an item as the maximum over all windows from the past until the current state. Hence, when the stream evolves, the length of the window containing the highest frequency for a given item can grow and shrink continuously. We show some important properties on how the length of the maximal window can evolve.

We study the properties of the new measure, and propose an incremental algorithm that allows to produce the current frequency of an item immediately at any time. Our algorithm maintains a small summary of relevant information of the history of the stream that allows to produce the current frequency of an item immediately at any time. That is, when a new item arrives, the summary is updated, and when at a certain point in time, the current frequency is required, the result can be obtained instantly from the summary. The structure of the summary is based on some critical observations about the windows with the maximal frequency. In short, many points in the stream can never become the starting point of a maximal window, no matter what the continuation of the stream will be. The summary will thus consist of some statistics about the few points in the stream that are still candidate starting points of a maximal window. These important points in the stream will be called the *borders*.

Critical for the usefulness of the technique are the memory requirements of the summary that needs to be maintained in memory. It is shown experimentally that the memory requirements of the algorithm are extremely small for many different realistic data distributions. Even though in worst case the summary depends on the length of the stream, for realistic data distributions, its size is extremely small. Obviously, this property is highly desirable as it allows for an efficient and effective computation of our new measure.

1 Ruoming J. and Agrawal G.: *An Algorithm for In-Core Frequent Itemset Mining on Streaming Data.* in Proc. 5th IEEE Int. Conf. on Data Mining (ICDM'05), pp 210–217.
2 Demaine E.D., Lopez-Ortiz A. and Munro, J.I.: *Frequency Estimation of Internet Packet Streams with Limited Space.* in Proc. of the 10th Annual European Symposium on Algorithms (2002), pp 348–360.
3 Giannella C., Han J., Robertson E. and Liu C.: *Mining Frequent Itemsets Over Arbitrary Time Intervals in Data Streams.* Technical Report TR587 at Indiana University, Bloomington, (Nov 2003), 37 pages.
4 Karp, R. M., Papadimitriou, C. H. and Shenker, S.: *A Simple Algorithm for Finding Frequent Elements in Streams and Bags.* in ACM Trans. on Database Systems (2003), 28, pp 51–55.