

What's Hot and What's Not: Tracking Most Frequent Items Dynamically

Review

This is another volatile area in the field of databases. This is a well-designed paper, containing a lot of mathematical analysis. The proposed technique for keeping track of frequently accessed items is explained in detail. The experiment is also explained in detail and the tradeoffs of the technique are also mentioned. The author mentioned what future work can be done in this area. Though it is a long paper to read but it is quite an insight of the work that has already been done in this area and also the currently developed techniques. The use synthetic as well as real data to analyze the results of the algorithms they proposed.

Facts learned

The initial work in this field was to find items which occurred more than half of the time. The recent research is in the area of processing data streams. The proposed algorithms focus on the top-k items in a distributed environment and the goal is to minimize communication. The common feature of the already done research is that they use counter. The problem with this approach is that this won't work in a dynamic environment where items are deleted and inserted at the same time. Another research work shows algorithms for maintaining histograms with guaranteed accuracy and small space.

A more recent work is of maintaining quantiles that keeps the sum of items in random subsets. This approach is similar to what the authors did in this paper. The algorithm proposed by the author is more efficient in terms of space and time as compared to existing ones. The literature talk about group testing and once the group is generated; we can maintain sums of deterministic sets given again by error correcting codes. The basic idea is that during insertion or deletion, this new algorithm takes the same action and does not inspect the contents of the memory. The counter is incremented or decremented as a function of the item value. The group testing guarantees that no infrequent items will be output. The counters are kept for subsets of items and the monitoring of items is fixed in advance. This approach helps in analyzing dynamic datasets and is also simple to implement.

Future research

The new algorithm allows aggregation of information from separate sources. The authors suggest contrasting this approach with other approaches mentioned in B. Babcock and C. Olston. Distributed top-k monitoring. In Proceedings of ACM SIGMOD, 2003. The new algorithm focuses on minimizing space and time but these approaches focus on minimizing time. The authors also mention that immediate comparison of the approaches is not possible but can be done for periodic updates.

Another area mentioned for research is to analyze difference in frequencies between difference datasets. This is helpful in areas like trend analysis, financial data sets and anomaly detection. The other areas for potential research are applying this approach

to designing summary data structures for maintenance of other statistics of interest and in data stream applications, finding combinatorial designs which can achieve the same properties as the randomly chosen subsets, in order to give a fully deterministic construction for maintaining frequently occurring items.

References from the paper

B. Boyer and J. Moore. A fast majority vote algorithm. Technical Report 35, Institute for Computer Science, University of Texas, 1982

M. Fischer and S. Salzberg. Finding a majority among n votes: Solution to problem 81-5. *Journal of Algorithms*, 3(4):376{379, 1982

J. Misra and D. Gries. Finding repeated elements. *Science of Computer Programming*, 2:143{152, 1982.

E. Demaine, A. Lopez-Ortiz, and J. I. Munro. Frequency estimation of internet packet streams with limited space. In *Proceedings of the 10th Annual European Symposium on Algorithms*, volume 2461 of *Lecture Notes in Computer Science*, pages 348-360, 2002.

R. Karp, C. Papadimitriou, and S. Shenker. A simple algorithm for finding frequent elements in sets and bags. *ACM Transactions on Database Systems*, 2003.

G. Manku and R. Motwani. Approximate frequency counts over data streams. In *Proceedings of 28th International Conference on Very Large Data Bases*, pages 346{357, 2002.

B. Babcock and C. Olston. Distributed top-k monitoring. In *Proceedings of ACM SIGMOD*, 2003.

P. Gibbons and Y. Matias. New sampling-based summary statistics for improving approximate query answers. In *Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD-98)*, volume 27 of *ACM SIGMOD Record*, pages 331 {342, 1998.

P. Gibbons and Y. Matias. Synopsis structures for massive data sets. *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, A, 1999.

P. B. Gibbons, Y. Matias, and V. Poosala. Fast incremental maintenance of approximate histograms. In *Proceedings of the 23rd International Conference on Very Large Data Bases (VLDB '97)*, pages 466{475, 1998.

A. Gilbert, S. Guha, P. Indyk, Y. Kotidis, S. Muthukrishnan, and M. Strauss. Fast, small-space algorithms for approximate histogram maintenance. In *Proceedings of the 34th ACM Symposium on Theory of Computing*, 2002.

A. C. Gilbert, Y. Kotidis, S. Muthukrishnan, and M. Strauss. How to summarize the universe: Dynamic maintenance of quantiles. In Proceedings of 28th International Conference on Very Large Data Bases, 2002.