Seminar on Advanced topics in data structures

Fall 2003/2004

Haim Kaplan
phone: 03-6408814
email: haimk@math.tau.ac.il

October 26, 2003

We shall focus on recent algorithms and data structures for processing large streams of data. Typically the model is that we see the elements of a huge steams one at a time, and we want to keep a concise summary (sometimes called synopsis) that allows to answer queries of some sort. Say as an example we may want a summary that allows to determine the number of distinct element in the stream. Typical data sets of interest are internet packet traces, market transactions, stock transactions etc. For recent surveys see the web site.

How to get in: To participate in the seminar you must choose a paper and register it with me by Thursday (30/10) of this week. Most of the paper are available online through links from the course web site. You can also get hard copies either by coming to my office or at the library. Once you decide that you want to participate and picked a paper please come to register the paper at my office (its also possible to do that by phone or email if you cannot come in person). Please note that as soon as you register your chances to get in with the paper that you want are higher, in particular if there are a lot of students in class. Papers in the list below are ordered more or less in the order in which I would like them to be presented so you can estimate the approximate week of your presentation.

Once registered you must do the presentation you committed to!

Your grade: You'd have to attend all meetings and actively participate in the discussion so as to make sure you understand the material. The papers are strongly related. You should follow the material covered before your presentation and connect your presentation to it. (That does not necessarily mean that you do not have to repeat crucial definitions etc.) You are encouraged to ask questions and you get credit for it. To encourage you to ask everything you like I guarantee not to take any credit from the speaker even in case that he or she does not answer well. So there is absolutely no way you can hurt your friend by asking, and you may hurt yourself if you refrain from asking.

Participation worth up to 15% of your final mark. The other 85% will be determined by the quality of your presentation.

Guidelines for preparing your presentation:

1. All papers on the list are mostly theoretical and you may have to spend some time in order

to understand your paper well. Good understanding of the paper is the key for a good presentation. So start working early and try to gain a good understanding of the results.

2. Some of the papers on the list are related to others. So attending the other lectures would certainly help you when you work on your paper.

3. You are not restricted to use any particular media for your presentation. Clear conventional blackboard talk certainly can gain maximum credit. Remember that sometimes a picture worth 1000 words. Media such as slides or computer slides is particularly useful when you want to show pictures and animate operations since doing that on the board may take too much time. Your time is limited so you have to use it carefully.

4. Sometimes with the time you have you will not be able to cover everything that is written in your paper. Depending on how crowded the seminar would be, it may not be possible to give you extra time. So plan the talk carefully, you need not cover everything in the paper but what you cover should be transfered clearly. Don't cover stuff that is in your paper but has already been covered by previous speakers.

   If you think that you need less time than what you have since the paper is simple or short, please let me know as soon as you estimate it.

References

[1] G. Manku and R. Motwani. Approximate frequency counts over data streams. In *Proceedings of the 28th International Conference on Very Large Data Bases*, 2002.

[2] M. Charikar, K. Chen, and M. Farach-Colton. Finding frequent items in data streams. In *Proceedings of 29th International Colloquium on Automata, Languages and Programming (ICALP)*, 2002.

[3] G. Cormode and S. Muthukrishnan. What's hot and what's not: Tracking most frequent items dynamically. In *Proceedings of the Twentieth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pages 296–307, 2003.

[4] Gurmeet Singh Manku, Sridhar Rajagopalan, and Bruce G. Lindsay. Approximate medians and other quantiles in one pass and with limited memory. In *Proceedings of the 1998 ACM SIGMOD international conference on Management of data*, pages 426–435. ACM Press, 1998.

[5] Michael Greenwald and Sanjeev Khanna. Space-efficient online computation of quantile summaries. In *Proceedings of the 2001 ACM SIGMOD international conference on Management of data*, pages 58–66. ACM Press, 2001.

[6] Anna C. Gilbert, Yannis Kotidis, S. Muthukrishnan, and Martin J. Strauss. How to summarize the universe: Dynamic maintenance of quantiles. In *In proceedings of the 28th Very Large Data Bases (VLDB) Conference*, 2002.

[7] Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. *Journal of Computer and Systems Sciences (JCSS)*, 58(1):137–147, 1999.

[8] P. Indyk. Stable distributions, pseudorandom generators, embeddings and data stream computation. In *Proceedings of the 41st IEEE Annual Symposium on Foundations of Computer Science (FOCS)*, pages 189–197, 2000.

[9] M. Datar, A. Gionis, P. Indyk, and R. Motwani. Maintaining stream statistics over sliding windows. *Siam J. on computing*, 31(6), 2002.

[10] Brain Babcock, Mayur Datar, Rajeev Motwani, and Liadan O'Callaghan. Maintaining variance and k-medians over data stream windows. In *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 234–243. ACM Press, 2003.

[11] Edith Cohen. Size-estimation framework with applications to transitive closure and reachability. *Journal of Computer and Systems Sciences (JCSS)*, 55(3):441–453, 1997.

[12] Bernard Chazelle, Joe Kilian, Ronitt Rubinfeld, and Ayellet Tal. The bloomier filter: An efficient data structure for static support lookup tables. In *ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2004.