CMSC 498K: Algorithms for analyzing massive data sets and data mining

Spring 2008

Course Information: *http://www.cs.umd.edu/~samir/498/*

Instructor: Samir Khuller - Office: AVW 3369. Office phone: (301) 405–6765.
E-mail: samir@cs.umd.edu.
Office hours: Tu Th 11–12 and by appointment.

Class Time: TuTh 9:30-10:45am. CSIC 3118.

Course Overview: As large amounts of data are being created it is important to understand how to analyse the data to extract interesting trends and patterns. This is called "Data mining". Since the volume of data is large, it may not be feasible to make more than a single pass over the data. Stream processing methods provide effective ways to extract useful information from large data sets by making very few passes on the data. Surprisingly, a lot of information can be gleaned by making a single pass over the data, or a small number of passes over the data. The first part of the course will cover random sampling and stream processing methods. We will also consider privacy issues in data bases and how these should be handled. The second part of the course will cover more traditional data mining topics such as clustering, statistical inference and association rule learning. Data warehousing and online analytical processing is another topic we will cover.

Material: We will cover material from a lot of papers. Some of the papers are actually not easily available online. An option would be to scan them in and put them up on the class page.

Papers on Stream Processing:

1. Selection and Sorting with limited storage (Theoretical Computer Science, 12:315–323, 1980) by J. Ian Munro and Mike Paterson
2. Finding a majority among N votes (J. Algorithms, 3: 375–379 1982) by Michael Fischer and Steven Salzberg
3. Probabilistic counting algorithms for data base applications (Journal of Computer and System Sciences, 31:182-209, 1985) by Phillipe Flajolet and Nigel Martin
4. Space Complexity of approximating frequency moments (Journal of Computer and System Sciences, 58:137–147, 1999) by Noga Alon, Yossi Matias and Mario Szegedy
5. A simpler and more efficient deterministic scheme for finding frequent items over sliding windows, Proc. ACM Principles of Database Conference (PODS) 2006, L.K. Lee, H.F. Ting
6. Incremenal Clustering and Dynamic Information retrieval Proc. of ACM Symp on Theory of Computing (STOC) 1997 Moses Charikar, Chandra Chekuri, Tomas Feder and Rajeev Motwani
7. Data streams and histograms Proc. of ACM Symp on Theory of Computing (STOC) 2001 by Sudipto Guha, Nick Koudas and Kyuseok Shim
8. Approximate Medians and Quantiles in One pass and with limited memory ACM SIGMOD Record, Vol 27:426–435, 1998 Gurmeet Manku, Sridhas Rajagopalan and Bruce Lindsay
9. Counting Triangles in Data Streams Proc. ACM Principles of Database Conference (PODS) 2006, Luciano Buriol, Gereon Frahling, Stefano Leonardi, Alberto Marchetti-Spaccamela, Christian Sohler
10. Survey paper by Richard Karp 2007 Foundations of Software Technology and Theoretical Computer Science Conference Proc. published by Springer Verlag as LNCS.

Papers on Privacy:

1. Achieving anonymity via clustering Proc. ACM Principles of Database Conference (PODS) 2006, Gagan Aggarwal, Tomas Feder, Krishnaram Kenthapadi, Samir Khuller, Rina Panigrahy, Dilys Thomas and An Zhu
2. Anonymizing tables Proc. International Conference on Database Theory 2005, Gagan Aggarwal, Tomas Feder, Krishnaram Kenthapadi, Rajeev Motwani, Rina Panigrahy, Dilys Thomas and An Zhu

Additional Readings: Principles of Data Mining by Hand, Mannila and Smyth (MIT Press, 2001).

Data Streams: Algorithms and Applications S. Muthukrishnan Now Publishers (2006)

J. Han and M. Kamber. Data Mining: Concepts and Techniques Morgan Kaufmann (2000).

MapReduce: Simplified Data Processing on Large Clusters Sixth Symposium on Operating System Design and Implementation (OSDI), 2004. Jeffrey Dean and Sanjay Ghemawat

Prerequisites: CMSC 351 or an equivalent course. I expect familiarity with basic algorithms. This course is an algorithmic oriented course, with proofs of correctness etc.

Course Work: Course work will consist of two exams (midterm and final). There will be a few homeworks during the semester. The graded homeworks must be done *by yourself* with no help from any outside sources. There will also be an in-class presentation. If you cannot come to class, please ask a friend to pick up copies of handouts from you. It is your responsibility to make sure that you pick up all handouts.

Grading: Grades will be based on the homeworks, and the two exams. The relative weights of these will be 20% for the homeworks, 10presentation, 30

Each of you should do the homework by yourself, without discussion with other students, or of use of the Internet to find solutions.