Course calendar

6.897 Algorithms for Massive Data Sets

November 11, 2002

1. *September 5:* Introduction. Computational models: main memory, external memory, streaming. Examples.

2. *September 10:* Numerical data (B-trees). Main memory algorithms for geometric data in 2D
   Reading:

   - B-trees: "Introduction to algorithms", Cormen, Leiserson, Rivest, Stein, 2nd edition.
   - K-d-trees: "Computational Geometry", M. de Berg, M. van Kreveld, M. Overmars, O. Szwarzkopf. Chapter 5.
   - Approximate Nearest Neighbor (general case, without the bounded ratio assumption): Arya et al, "An Optimal Algorithm for Approximate Nearest Neighbor Searching", Journal of the ACM (1998), 891-923. For implementation, see `http://www.cs.umd.edu/~mount/ANN`

3. *September 12:* Low dimension ctd.. External memory algorithms for orthogonal line intersection.
   Reading: see `http://www.brics/dk/MassiveData02`, especially the 2nd lecture by Jeff Vitter.

4. *September 17:* Geometric data in high dimensions. Dimensionality reduction in Hamming cube.
   Reading: A very similar (with minor differences) dimensionality reduction technique is described in my Ph.D. thesis, section 4.1 (available at my web page). See the references in there for other (earlier) approaches.

5. *September 19:* High dimensions ctd. Continuation of Hamming cube, and Min Hashing.
   Reading: Min Hashing was introduced in Broder et al, WWW6, 1997.

6. *September 24:* High dimensions ctd. Details on Min Hashing. Near neighbor search via Locality Sensitive Hashing. Applications to web clustering.
   Reading:

- Linear permuations: Broder, Charikar, Frieze, Mitzenmacher, "Min-wise independent permutations", STOC'98.
  Available at `http://www.cs.princeton.edu/~moses/papers/`
- Polynomial functions: Indyk, "A small approximately min-wise independent family of hash functions", SODA'99. Available at my web page.
- Locality Sensitive Hashing: Indyk,Motwani, "Approximate nearest neighbor: towards removing the curse of dimensionality", STOC'98, section 4.2. Note: "Near Neighbor" is called "PLEB" in that paper.
- Web clustering I: Broder et al, WWW6, 1997.
- Web clustering II: Haveliwala, Gionis, Indyk, "Scalable Techniques for Clustering the Web", WebDB 2000. Available at my web page.

7. *September 26:* High dimensions ctd. Dimensionality reduction.

   - Median estimator for l_p norms: P. Indyk, "Stable Distributions, Pseudorandom Generators, Embeddings and Data Stream Computation", FOCS 2000. Available at my web page.
   - Norm estimator for l_2, i.e., Johnson-Lindenstrauss lemma:
     - Original proof: W.B. Johnson and J. Lindenstrauss, "Extensions of Lipshitz mapping into Hilbert space", Contemporary Mathematics (26), 1984, p. 189-206.
     - Simplification of the original proof: P. Frankl and H. Maehara, "The Johnson-Lindenstrauss Lemma and the Sphericity of Some Graphs", Journal of Combinatorial Theory B (44), 1988, p. 355-362.
     - Gaussian-based variant of the proof: P. Indyk and R. Motwani, "Approximate Nearest Neighbors: Towards Removing the Curse of Dimensionality", 30th Symposium on Theory of Computing , 1998 (appendix). Available at my web page.

8. *October 1:* High dimensions ctd. Singular Value Decomposition (iterative methods, sampling)
   Reading:

   - General SVD and power method: G. H. Golub and C. F. Van Loan, "Matrix Computations".
   - Sampling-based algorithm: Petros Drineas, Ravi Kannan, Alan Frieze, Santosh Vempala and V. Vinay, "Clustering in large graphs and matrices", Proc. of the 10th ACM-SIAM Symposium on Discrete Algorithms, Baltimore, 1999.

9. *October 3:* Clustering. k-means, EM, agglomerative methods.
   Reading:

   - EM: A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm", Journal of the Royal Statistical Society, Series B (Methodological) , 39(1):1–38, 1977. This is the most cited Computer Science article, according to CiteSeer, see `http://citeseer.nj.nec.com/articles.html` for more info. Note that some *books* are cited more often.

- Agglomerative Hierarchical Clustering: type "agglomerative hierarchical clustering" in Google and see what happens.

10. *October 8:* Clustering ctd. Sampling-based algorithms.
    Reading:

    - Sampling-based algorithm for k-median: folklore. A more complex approach (clustering large clusters, identifying points from small clusters, and re-clustering those points) appeared in P. Indyk, "Sublinear-time algorithms for metric space problems", STOC'99.
    - Greedy algorithm: an adaptation of an algorithm by M. Charikar, O'Callaghan and R. Panigrahy.

11. *October 10:* Clustering ctd.
    Reading: S. Guha, N. Mishra, R. Motwani, and L. O'callahan, "Clustering Data Streams", available at `http://www.cis.upenn.edu/ sudipto/papers.html`

12. *October 15:* **Holiday - no classes**

13. *October 17:* Streaming. Norm estimation.
    Reading: same as for Lecture 7

14. *October 22:* Streaming ctd. Approximate representation of data.
    Reading: the paper is available at my web page.

15. *October 24:* Overview of PS1

16. *October 29:* Text data. Inverted index, suffix trees.
    Reading: the slides by Paolo Ferragina at `http://www.brics.dk/MassiveData02/`.
    Note: the PDF file seems to show nonsense, use PPT or PS files.

17. *October 31:* Text ctd. Space efficient suffix trees.
    Reading: as above.

18. *November 5:* Graph data. Spectral partitioning, finding dense subgraphs.
    Reading: to be determined, but the paper "Spectral partitioning of random graphs" by F. McSherry is a good approximation.
    See `http://www.cs.washington.edu/homes/mcsherry/spectral.html`.

19. *November 7:* Graphs ctd. Models for massive graphs.
    Reading: to be determined, but the paper "Stochastic models for the Web graph", by Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, D. Sivakumar, Andrew Tomkins, Eli Upfal is a good approximation.
    See `http://citeseer.nj.nec.com/kumar00stochastic.html`

20. *November 12:* Graphs ctd. Link analysis (Pagerank etc).
    Reading: most likely "Spectral analysis for data mining", by Y. Azar, A. Fiat, A. Karlin, F. McSherry and J. Saia.
    See `http://www.cs.washington.edu/homes/mcsherry/spectral.html`.

21. *November 14:* Compression. Huffman coding, Lempel-Ziv.
    Reading: For Huffman codes see "Introduction to Algorithms" by Cormen, Leiserson, Rivest, Stein, p. 385.

22. *November 19:* 2D Compression.

23. *November 21:* Erasure-correcting-codes.
    Reading: Basic information on codes can be found in Madhu's notes at `http://theory.lcs.mit.edu/~madhu/coding/course.html`. In the first lecture we will cover Reed-Solomon codes: definition, construction, proof of large distance, and encoding + decoding from *erasures*.

24. *November 26:* Codes ctd.
    Reading: "Efficient Erasure Correcting Codes", Michael G. Luby, Michael Mitzenmacher, M. Amin Shokrollahi and Daniel A. Spielman, at `http://www-math.mit.edu/~spielman/Research/erase.html`. Note: the paper contains lots of fairly complex math; for the presentation, I will either try to find a simpler proof, or we will skip the more complex math altogether.

25. *November 28:* **Holiday - no classes**.

26. *December 3:* Exciting topic II

27. *December 5:* Exciting topic III

28. *December 10:* Summary.