Randomized Algorithms for Matrices and Massive Data Sets

Petros Drineas
Dept. of Computer Science
Rensselaer Polytechnic Institute
Troy, NY 12180
drinep@cs.rpi.edu

Michael W. Mahoney
Yahoo Research Labs
Sunnyvale, CA 94089
mahoney@yahoo-inc.com

ABSTRACT

The tutorial will cover randomized sampling algorithms that extract structure from very large data sets modeled as matrices or tensors. Both provable algorithmic results and recent work on applying these methods to large biological and internet data sets will be discussed.

1. TUTORIAL SUMMARY

Large matrices arise in numerous applications. For example, in Information Retrieval, Data Management, and Data Mining, the data often consist of m objects, e.g., documents, genomes, images, or web pages, each of which may be described by n features. Such data may be represented by an $m \times n$ matrix A, the rows of which are the object vectors and the columns of which are the feature vectors. Similarly, large tensors arise in applications in which the data are described by a variable subscripted by three or more indices.

The tutorial will cover recent advances in theoretical techniques for handling large data sets in the form of matrices or tensors, including both theoretical foundations of these techniques and their applications in the context of VLDB research topics. Applications of these techniques include explaining the success of LSI for large document corpora, nearest neighbor queries, "sketching approaches" for matrices and tensors, speeding up kernel computations, recommendation systems and collaborative filtering, and a large class of algorithmic problems in the framework of the streaming and the pass-efficient model.

We will cover:

- provably accurate sampling based algorithms for computing: the product of two or more matrices; the SVD of a matrix; CUR decompositions of a matrix,
- theoretical limitations of these techniques,
- applications of such algorithms in traditional data mining problems such as nearest neighbor queries, recommendation systems, and speeding up of kernel computations in machine learning,
- empirical evaluation of such approaches in the context of bioinformatics and medical datasets,
- empirical evaluation of such approaches in internet and recommendation system applications, and
- methods to extend the presented results and develop new algorithms by coupling them with other more traditional tools of data analysis.

The presentation will be available at our web pages (for Drineas: `http://www.cs.rpi.edu/~ drinep`; and for Mahoney: `http://www.cs.yale.edu/homes/mmahoney/`). In addition, please see:

- `http://www.cs.yale.edu/homes/mmahoney/talks/KDD05_dm.ppt` (for a similar tutorial given at the 2005 ACM SIGKDD conference), and
- `http://www.cs.yale.edu/homes/mmahoney/talks/SDM06_dm.ppt` (for a similar tutorial given at the 2006 SIAM Data Mining conference).

For more details about some of the material to be presented, see [1, 2, 3, 4, 5, 6].

2. REFERENCES

[1] P. Drineas, R. Kannan, and M.W. Mahoney. Fast Monte Carlo algorithms for matrices I: Approximating matrix multiplication. *SIAM Journal on Computing*, 36:132–157, 2006.

[2] P. Drineas, R. Kannan, and M.W. Mahoney. Fast Monte Carlo algorithms for matrices II: Computing a low-rank approximation to a matrix. *SIAM Journal on Computing*, 36:158–183, 2006.

[3] P. Drineas, R. Kannan, and M.W. Mahoney. Fast Monte Carlo algorithms for matrices III: Computing a compressed approximate matrix decomposition. *SIAM Journal on Computing*, 36:184–206, 2006.

[4] P. Drineas, M.W. Mahoney, and S. Muthukrishnan. Polynomial time algorithm for column-row based relative-error low-rank matrix approximation. Technical Report 2006-04, DIMACS, March 2006.

[5] M.W. Mahoney, M. Maggioni, and P. Drineas. Tensor-CUR decompositions for tensor-based data. In *Proceedings of the 12th Annual ACM SIGKDD Conference*, 2006.

[6] P. Paschou, M.W. Mahoney, J.R. Kidd, A.J. Pakstis, S. Gu, K.K. Kidd, and P. Drineas. Intra- and inter-population genotype reconstruction from tagging SNPs. *Manuscript submitted for publication.*

Permission to copy without fee all or part of this material is granted provided that the copies are not made or distributed for direct commercial advantage, the VLDB copyright notice and the title of the publication and its date appear, and notice is given that copying is by permission of the Very Large Data Base Endowment. To copy otherwise, or to republish, to post on servers or to redistribute to lists, requires a fee and/or special permission from the publisher, ACM.

VLDB'06, September 12–15, 2006, Seoul, Korea.

Copyright 2006 VLDB Endowment, ACM 1-59593-385-9/06/09.