Report on Dagstuhl seminar
"Web Information Retrieval and Linear Algebra Algorithms"
11 February 2007 – 16 February 2007

organized by

Andreas Frommer, Department of Mathematics and Computer Science, Wuppertal
Michael W. Mahoney, Yahoo Research, USA
Daniel B. Szyld, Department of Mathematics, Temple University, Philadelphia, USA.

Abstract

A seminar concentrating on the intersection of the fields of information retrieval and other web-related aspects with numerical and applied linear algebra techniques was held with the attendance of scientists from industry and academia.

1 Goals of the seminar

The scientific community has witnessed the increasing importance of linear algebra algorithms and of Markov chain modeling in several applications from computer science. Of particular importance is linear algebra algorithms to study the structure of the Web and information retrieval (IR) on the Web. The main focus of the seminar was the evolving theory and computational aspects of methods for web information retrieval, including search engines, that are inspired by traditional and recent advances in algorithms for linear algebra problems. To this end, the seminar brought together scientists from academia with background in computer science or numerical mathematics and scientists working in industry, mostly from Yahoo Research (both from the US and Europe).

Dagstuhl Seminar Proceedings 07071
Web Information Retrieval and Linear Algebra Algorithms
http://drops.dagstuhl.de/opus/volltexte/2007/1070

2 Structure of the seminar

The seminar was attended by forty-seven participants coming from thirteen different countries. We had a good mixture of graduate students, young researchers, scientists in mid-career, and senior investigators from academia and industry. There was a total of thirty-one talks. Due to the diverse backgrounds of the attendees it was decided to have five longer expository talks which included introductions to the subjects and methods of the respective fields. These 'tutorials' were:

- Pagerank acceleration and sensitivity analysis (Chen Greif)
- Iteration at different levels - on multi level approaches for computing the stationary distribution of large Markov chains (Peter Buchholz)
- Using non-negative matrix and tensor factorizations for email surveillance (Michael Berry)
- Sampling algorithms for matrices and data (Michael Mahoney)
- Web term dependence issues in document retrieval (Hugo Zaragoza)

The other talks were scheduled in thematic sessions with substantial time reserved for discussions and interactions.

3 Outcome of the seminar

We want to highlight that the seminar really fostered interaction between people from academia and industry. Many participants observed that they benefited greatly from the contributions presented from researchers working in other fields or other settings.

Among the findings of this seminar, we mention the following: While it became clear from the scientists working in web retrieval that Pagerank now is just a minor ingredient in web ranking algorithms, it turns out that Pagerank-like approaches continue to play an important role in other areas such as social science or community behavior. In this area, but also in more advanced, semantic models, the properties of eigenvalues and eigenvectors of huge sparse matrices and their computation continue to be at the heart of current research. Similarly, other classical matrix factorization techniques like the singular value decomposition have new applications, for example, in cluster analysis.

Techniques using low rank (and thus data efficient) approximations to huge matrices become increasingly important for data analysis and representation. For example, recent work has focused on employing randomization to improve low-rank computations and also large statistical regression problems. A particularly difficult issue is that traditional methods such as the SVD and QR decomposition destroy sparsity. Thus, low-rank approximations that respect sparsity are important. A second issue is that in many applications, one is not interested in the results of low-rank computations per se, but instead one wants to use it to learn from the data. Thus, studying matrix decompositions with good learning or generalization properties is important. Relatedly, in many cases an important question has to do with the best way to represent the data, i.e., which vector space is most appropriate to model the data in order to perform efficient computations.

Asynchronous iterative approaches, as they arise naturally in loosely coupled networks of processors have been analyzed from the theoretical side and are being used in practice. One challenging problem discussed, was that of data streams which cannot be stored, so that standard numerical techniques have to be enhanced, for example, with statistical analyses or using novel algorithmic methods. Another point of intersection between the disciplines were novel graph partitioning approaches using iterative methods from numerical linear algebra. This represents a particularly challenging direction since the local geometry of the data that arise in Web IR applications is very different than the geometry that arises in traditional applications.