07071 Abstracts Collection
Web Information Retrieval and Linear Algebra Algorithms
— Dagstuhl Seminar —

Andrea Frommer[1], Michael W. Mahoney[2] and Daniel B. Szyld[3]

[1] Univ. Wuppertal, DE
Andreas.Frommer@math.uni-wuppertal.de
[2] Yahoo Research - Sunnyvale, US
mahoney@yahoo-inc.com
[3] Temple Univ. - Philadelphia, US
szyld@temple.edu

Abstract. From 12th to 16th February 2007, the Dagstuhl Seminar 07071 "Web Information Retrieval and Linear Algebra Algorithms" was held in the International Conference and Research Center (IBFI), Schloss Dagstuhl. During the seminar, several participants presented their current research, and ongoing work and open problems were discussed. Abstracts of the presentations given during the seminar as well as abstracts of seminar results and ideas are put together in this paper. The first section describes the seminar topics and goals in general. Links to extended abstracts or full papers are provided, if available.

Keywords. Information retrieval, Markov chains, PageRank, numerical linear algebra, low rank approximations, sparsity, ranking

07071 Report on Dagstuhl seminar – Web Information Retrieval and Linear Algebra Algorithms

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).

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.

Joint work of: Frommer, Andreas; Mahoney, Michael W.; Szyld, Daniel B.

Extended Abstract: http://drops.dagstuhl.de/opus/volltexte/2007/1070

Why Spectral Retrieval Works (and when it does not)

Holger Bast (MPI für Informatik - Saarbrücken, D)

Spectral retrieval is a popular approach to ranked retrieval on large document collections. The basic scheme is as follows. Assumed that documents as well as queries are represented as high-dimensional vectors, where the dimensions correspond to the distinct words in the collection. The documents and the query are then projected on a low-dimensional eigenspace computed from the document vectors. Documents are then ranked according to their similarity to the query in that eigenspace.

Previous explanations for why spectral retrieval works are all of the following kind: if there are k "base" documents such that each document can be approximated by a linear combination of these, then reduction to an eigenspace of dimension k works. This kind of explanation leaves open the central question, as to what an appropriate choice for k is.

We argue that makes spectral retrieval work in practice is its ability to identify pairs of terms with similar co-occurence patterns. We give a parameterless algorithm that on a number of test collections outperforms all previous algorithms committing to a fixed dimension. We also mention a recent extension of our approach that is able to indentify asymmetric relations between terms (banana - fruit).

Keywords: Spectral retrieval, dimension reduction, term taxonomy, search engine

Joint work of: Bast, Holger; Majumdar, Debapriyo

Using Non-negative Matrix and Tensor Factorizations for Email Surveillance

Michael Berry (University of Tennessee, USA)

Automated approaches for the identification and clustering of semantic features or topics are highly desired for text mining applications. Using a low rank non-negative matrix factorization (NNMF) algorithm to retain natural data non-negativity, we eliminate the need to use subtractive basis vector and encoding calculations present in techniques such as principal component analysis for semantic feature abstraction. Using non-negative tensor factorization (NNTF), temporal proximity can be exploited to enable tracking of focused discussions.

Demonstrations of NNMF and NNTF algorithms for topic (or discussion) detection and tracking using the Enron Email Collection is presented.

Keywords: Non-negative matrix and tensor factorization, email surveillance, Enron email collection

Joint work of: Berry, Michael W. ; Bader, Brett W. ; Browne, Murray

Schwarz Iterations for PageRank Matrices

Stefan Borovac (Universität Wuppertal, D)

We give a an overview on a graph based convergence theory for PageRank matrices which includes the convergence of additive and multiplicative Schwarz.

It will be discussed under which conditions a state space reordering is necessary to achieve convergence and whether the overlap has an impact to the convergence or not.

PageRank matrices are normally irreducible and have a positive diagonal. Thus, it turns out that reordering is not necessary and that the overlap can be chosen with a reliable degree of freedom if additive Schwarz iterations are used. In the case of multiplicative Schwarz there are certain restrictions which must be considered to achieve convergence.

The theory will be discussed for both, one-level and two-stage Schwarz iterations and for standard iteration operators as well as shifted (relaxed) iteration operators.

Keywords: Linear systems, PageRank, Markov chains, singular matrices, iterative methods, block methods, additive Schwarz, multipl icative Schwarz, matrix graph theory

Iteration at Different Levels: Multi-Level Methods fro Structured Markov Chains

Peter Buchholz (Universität Dortmund, D)

For the stationary analysis of large Markov chains in continuous and discrete time a wide variety of solution techniques has been applied in the past. Empirical comparisons show that in particular so called multi-level approaches that perform iterations at different levels are the most efficient solvers for a wide class of Markov chains. The methods combine ideas from aggregation disaggregation methods and algebraic multigrid.

The talk gives an overview of the basic ideas of multi level approaches and shows which design alternatives for the algorithms exist. In particular it considers different forms of defining levels, available alternatives to realize prolongation and interpolation operations, different cycle types and different stopping criteria for the smoothing operations at each level. The last part of the talk is devoted to implementation issues and data structures that are necessary for an efficient realization of multi-level methods.

Keywords: Stationary analysis, multi-Level techniques, Kronecker representation

Full Paper: http://drops.dagstuhl.de/opus/volltexte/2007/1059

Steady-state analysis of Google-like matrices with structured methods

Tugrul Dayar (Bilkent University - Ankara, TR)

The mathematical tool of Markov chains (MCs) used in performance modeling emerges in many research areas today. The dynamically changing matrices used by the Google search engine in ranking documents are among the largest of these. These matrices are not only large, but they are also sparse, reducible stochastic matrices with some zero rows. Ranking documents amounts to solving for the positive left-hand eigenvectors of linear combinations of these matrices with appropriately chosen rank-1 matrices. The most suitable method of choice for this task appears to be the power method. Various improvements have been obtained using techniques such as quadratic extrapolation and aggregation-disaggregation. In this talk, we show that another kind of improvement may be obtained by using ideas from graph theory and computing cutsets for the state partitions defining the irreducible diagonal blocks of the underlying reducible, sparse matrices. Since this approach can also be used with other kinds of matrices including irreducible ones, it seems to have a large application area.

Keywords: Google-like matrices, cutsets

Joint work of: Dayar, Tugrul; Noyan, Gokce Nil

Ranking a stream of News

Gianna Maria Del Corso (Università di Pisa, I)

According to a recent survey made by Nielsen NetRatings, searching on news articles is one of the most important activity online. Indeed, Google, Yahoo, MSN and many others have proposed commercial search engines for indexing news feeds. Despite this commercial interest, no academic research has focused on ranking a stream of news articles and a set of news sources. In this paper, we introduce this problem by proposing a ranking framework which models: (1) the process of generation of a stream of news articles, (2) the news articles clustering by topics, and (3) the evolution of news story over the time. The ranking algorithm proposed ranks news information, finding the most authoritative news sources and identifying the most interesting events in the different categories to which news article belongs. All these ranking measures take in account the time and can be obtained without a predefined sliding window of observation over the stream. The complexity of our algorithm is linear in the number of pieces of news still under consideration at the time of a new posting. This allow a continuous on-line process of ranking. Our ranking framework is validated on a collection of more than 300,000 pieces of news, produced in two months by more then 2000 news sources belonging to 13 different categories (World, U.S, Europe, Sports, Business, etc). This collection is extracted from the index of COMETOMYHEAD, an academic news search engine available online.

Keywords: News engines, information extraction, news ranking

Joint work of: Del Corso, Gianna Maria; Gulli', Antonio; Romani, Francesco;

See also: G. M. Del Corso, A. Gulli', F. Romani, Ranking a Stream of News, In Proc. of the 14th Int. WWW Conf., Chiba, Japan, 2005. ACM Press, pp. 97–106.

Restarted Krylov subspace methods for the transient solution of Markov processes

Michael Eiermann (TU Bergakademie Freiberg, D)

The state probability vector of a homogeneous time-continuous Markov chain (with finitely many states) can be computed by evaluating the matrix exponential of the transition rate matrix times the initial probability distribution. We apply a recently developed restarted Arnoldi method to this problem. We derive two mathematically equivalent formulations of this restarted algorithm, the second of which, while slightly more expensive, was found to be more stable in the presence of rounding errors. We further present a-posteriori lower and upper bounds for the approximation error which lead to a reliable stopping criterion.

Finally, we demonstrate the performance of the restarted algorithm for two test problems.

Keywords: Markov chain, matrix exponential function, restarted Arnoldi method

Joint work of: Afanasjew, Martin; Eiermann, Michael; Ernst, Oliver; Güttel, Stefan

On Information Dissemination in Large Networks

Robert Elsässer (Universität Paderborn, D)

One frequently studied problem in the context of information dissemination in communication networks is the broadcasting problem. In this talk we consider the following randomized broadcasting protocol: At some time t an information r is placed on one of the nodes of a graph G. In the succeeding steps, each node chooses one neighbor, independently and uniformly at random, and opens a communication chanel to it. Then, any informed node sends a copy of r over each incident communication chanel opened in this step.

The talk consists of two parts. In the first part we state a strong relationship between randomized broadcasting and simple diffusion schemes. In particular, we show that the runtime of the algorithm described above is upper bounded by the runtime of simple diffusion schemes, up to a logarithmic factor. One key ingredient of our proofs is the analysis of a continous type version of the afore mentioned algorithm, which might be of independent interest.

In the second part we consider the communication overhead produced by the broadcasting algorithm described above in random-like graphs. We show that an apparently minor change in the ability of the nodes implies an exponential decrease in the average communication overhead produced by an arbitrary node. More precisely, we prove that if the nodes are allowed to remember the addresses of the neighbors chosen in the most recent three steps, then the communication overhead decreases substantially.

Keywords: Randomized broadcasting, diffusion schemes, random graphs

Joint work of: Elsässer, Robert; Sauerwald, Thomas

Multidamping simulation framework for link-based ranking

Efstratios Gallopoulos (University of Patras, GR)

We review methods for the approximate computation of PageRank. Standard methods are based on the eigenvector and linear system characterizations. Our starting point are recent methods based on series representation whose coefficients are damping functions, for example Linear Rank, HyperRank and TotalRank, etc. We propose a multidamping framework for interpreting PageRank

and these methods. Multidamping is based on some new useful properties of Google type matrices. The approach can be generalized and could help in the exploration of new approximations for list-based ranking. This is joint work with Georgios Kollias and is supported by a Pythagoras-EPEAEK-II grant.

Keywords: PageRank, Google, power method, eigenvalues, teleportation, list-based ranking, TotalRank

Joint work of: Georgios Kollias; Gallopoulos, Efstratios

Full Paper: http://drops.dagstuhl.de/opus/volltexte/2007/1060

Some results on the sensitivity of the PageRank vector

David Gleich (Stanford University, USA)

The PageRank Markov chain modification creates a new Markov chain with a unique stationary probability distribution from any finite state discrete time Markov chain. In this talk, we discuss the eigenstructure of the new transition matrix with some possible sensitivity implications. The main result is that the eigenvectors of the PageRank Markov chain are simple transformations of the eigenvectors of the original Markov chain, with the exception of the new dominant eigenvector.

The next results concern the derivative of the PageRank vector as a function of the mixing parameter α. We show that taking a Taylor step along the derivative is equivalent with computing a PageRank vector for a different teleportation distribution and we show the relationship between the derivation of the derivative from both a linear system and eigensystem formulation of the PageRank equation.

We illustrate additional properties of the PageRank derivative through a series of numerical examples

Keywords: PageRank, sensitivity, Markov chains, derivative

Joint work of: Gleich, David; Glynn, Peter; Golub, Gene; Greif, Chen

Three results on the PageRank vector: eigenstructure, sensitivity, and the derivative

David Gleich (Stanford University, USA)

The three results on the PageRank vector are preliminary but shed light on the eigenstructure of a PageRank modified Markov chain and what happens when changing the teleportation parameter in the PageRank model.

Computations with the derivative of the PageRank vector with respect to the teleportation parameter show predictive ability and identify an interesting set of pages from Wikipedia.

Keywords: PageRank, PageRank derivative, PageRank sensitivity, PageRank eigenstructure

Joint work of: Gleich, David; Glynn, Peter; Golub, Gene; Greif, Chen

Full Paper: http://drops.dagstuhl.de/opus/volltexte/2007/1061

PageRank Sensitivity and the Eigenstructure of the PageRank Markov Chain Transition Matrix

Gene H. Golub (Stanford University, USA)

The PageRank Markov chain modification creates a new Markov chain with a unique stationary probability distribution from any finite state discrete time Markov chain. In this talk, we discuss the eigenstructure of the new transition matrix with some possible algorithmic implications. The main result is that the eigenvectors of the PageRank Markov chain are simple transformations of the eigenvectors of the original Markov chain, with the exception of the new dominant eigenvector.

The second set of results concerns the derivative of the PageRank vector as a function of the mixing parameter α. We show the relationship between the derivation of the derivative from both a linear system and eigensystem formulation for the PageRank equation. These results lead to a set of problems related to computing *stable* PageRank distributions with respect to the mixing parameter α.

We will illustrate these results through a series of numerical examples.

Keywords: PageRank, sensitivity, Markov chains, derivative

Joint work of: Gleich, David F.; Glynn, Peter; Golub, Gene H.; Grief, Chen

PageRank Acceleration and Sensitivity Analysis

Chen Greif (University of British Columbia - Vancouver, CA)

In the first part of my talk, I will present a new stationary iterative scheme for PageRank computation. The algorithm is based on a linear system formulation of the problem and utilizes inner/outer iterations. It can be easily implemented and parallelized, and involves a minimal overhead beyond matrix-vector products. Convergence analysis is performed and it is shown that the algorithm is effective for a crude inner tolerance.

In the second part of the talk I will discuss new results on the rate of change of the first derivative of the stationary distribution vector of a rank-1 perturbation of a finite Markov chain. A measure of sensitivity is defined, which may be useful for determining the sensitivity of the PageRank vector to changes in the damping factor. The choice of the norm for computing the derivatives plays an

important role and affects the numerical procedure that is applied for evaluating the sensitivity measure.

The first part of the talk describes joint work with Andrew Gray and Tracy Lau, and the second part is joint work with Joel Friedman.

Keywords: PageRank acceleration, inner/outer iterations, sensitivity analysis

An Inner/Outer Stationary Iteration for Computing PageRank

Chen Greif (University of British Columbia - Vancouver, CA)

We present a stationary iterative scheme for PageRank computation. The algorithm is based on a linear system formulation of the problem, uses inner/outer iterations, and amounts to a simple preconditioning technique. It is simple, can be easily implemented and parallelized, and requires minimal storage overhead. Convergence analysis shows that the algorithm is effective for a crude inner tolerance and is not particularly sensitive to the choice of the parameters involved. Numerical examples featuring matrices of dimensions up to approximately 10^7 confirm the analytical results and demonstrate the accelerated convergence of the algorithm compared to the power method.

Keywords: PageRank, power method, stationary method, inner/outer iterations, damping factor

Joint work of: Gray, Andrew P.; Greif, Chen; Lau, Tracy

Full Paper: http://drops.dagstuhl.de/opus/volltexte/2007/1062

Nonlinear Approximation and Image Representation using Wavelets

Boulos Harb (Univ. of Pennsylvania, USA)

We address the problem of finding sparse wavelet representations of high-dimensional vectors. We present a lower-bounding technique and use it to develop an algorithm for computing provably-approximate instance-specific representations minimizing general ℓ_p distances under a wide variety of compactly-supported wavelet bases. More specifically, given a vector $f \in \mathbb{R}^n$, a compactly-supported wavelet basis, a sparsity constraint $B \in \mathbb{Z}$, and $p \in [1, \infty]$, our algorithm returns a B-term representation (a linear combination of B vectors from the given basis) whose ℓ_p distance from f is a $O(\log n)$ factor away from that of the optimal such representation of f. Our algorithm applies in the one-pass sublinear-space data streaming model of computation, and it generalize to weighted p-norms and multidimensional signals. Our technique also generalizes to a version of the problem where we are given a bit-budget rather than a term-budget. Furthermore, we use it to construct a *universal representation* that consists of at most $B(\log n)^2$ terms and gives a $O(\log n)$-approximation under all p-norms simultaneously.

Keywords: Nonlinear approximation, wavelets, approximation algorithms, streaming algorithms

Joint work of: Guha, Sudipto; Harb, Boulos

Full Paper: http://drops.dagstuhl.de/opus/volltexte/2007/1063

Full Paper: http://portal.acm.org/citation.cfm?id=1109633

The Sinkhorn-Knopp Algorithm: Convergence and Applications

Philip Knight (The University of Strathclyde - Glasgow, GB)

As long as a square nonnegative matrix A contains sufficient nonzero elements, the Sinkhorn-Knopp algorithm can be used to balance the matrix, that is, to find a diagonal scaling of A that is doubly stochastic.

We relate balancing to problems in traffic flow and describe how balancing algorithms can be used to give a two sided measure of nodes in a graph. We show that with an appropriate modification, the Sinkhorn-Knopp algorithm is a natural candidate for computing the measure on enormous data sets.

Keywords: Matrix balancing, Sinkhorn-Knopp algorithm, PageRank, doubly stochastic matrix

Full Paper: http://drops.dagstuhl.de/opus/volltexte/2007/1064

A Fast Algorithm for Matrix Balancing

Philip Knight (The University of Strathclyde - Glasgow, GB)

As long as a square nonnegative matrix A contains sufficient nonzero elements, then the matrix can be balanced, that is we can find a diagonal scaling of A that is doubly stochastic. A number of algorithms have been proposed to achieve the balancing, the most well known of these being the Sinkhorn-Knopp algorithm. In this paper we derive new algorithms based on inner-outer iteration schemes. We show that the Sinkhorn-Knopp algorithm belongs to this family, but other members can converge much more quickly. In particular, we show that while stationary iterative methods offer little or no improvement in many cases, a scheme using a preconditioned conjugate gradient method as the inner iteration can give quadratic convergence at low cost.

Keywords: Matrix balancing, Sinkhorn-Knopp algorithm, doubly stochastic matrix, conjugate gradient iteration

Joint work of: Knight, Philip A.; Ruiz, Daniel

Full Paper: http://drops.dagstuhl.de/opus/volltexte/2007/1073

Asynchronous PageRank computation in an interactive multithreading environment

Giorgios Kollias (University of Patras, GR)

Numerical Linear Algebra has become almost indispensable in Web Information Retrieval.

In this presentation we suggest that the asynchronous computation model is an attractive paradigm for organizing concurrent computations spanning data on Web scale. This suggestion is supported by experiments which highlight some interesting characteristics of this model as applied to 'page ranking' methods.

After an introduction on asynchronous computing in general and 'page ranking' in particular, we present results from the asynchronous compution of PageRank using typical combinations of execution units (processes, threads) and communication mechanisms (message passing, shared memory). Sound convergence properties predicted by theory are numerically verified and interesting patterns of behavior are unveiled. Our experiments were performed on Jylab, an evolving environment enabling interactive multithreading and multiprocessing computations. This work is supported by a Pythagoras-EPEAEK-II grant and is conducted in collaboration with Daniel Szyld.

Keywords: Asynchronous, pagerank, multithreading, multiprocessing

Joint work of: Kollias, Giorgios; Gallopoulos, Efstratios

Full Paper: http://drops.dagstuhl.de/opus/volltexte/2007/1065

Clustering with Matrix Decompositions

Amy N. Langville (College of Charleston, USA)

Matrix decomposition methods can be used to uncover clusters hidden in a large data matrix. When the matrix is square, typically spectral clustering methods are used. Often when the matrix is rectangular, the new nonnegative matrix factorization is used instead. In this talk, we extend spectral clustering methods to rectangular matrices by use of the sign patterns in the singular value decomposition. In order to compare the various clustering methods, all methods will be applied to the large familiar datasets derived from Reuters news feeds and the Enron email collection. If time permits, these clustering results will be related to the rank aggregation problem.

Keywords: Spectral clustering, singular value decomposition, nonnegative matrix factorization

Exploiting Community Behavior for Enhanced Link Analysis and Web Search

Julia Luxenburger (MPI für Informatik - Saarbrücken, D)

Methods for Web link analysis and authority ranking such as PageRank are based on the assumption that a user endorses a Web page when creating a hyperlink to this page. There is a wealth of additional user-behavior information that could be considered for improving authority analysis, for example, the history of queries that a user community posed to a search engine over an extended time period, or observations about which query-result pages were clicked on and which ones were not clicked on after a user saw the summary snippets of the top-10 results.

We study enhancements of link analysis methods by incorporating additional user assessments based on query logs and click streams, including negative feedback when a query-result page does not satisfy the user demand or is even perceived as spam. Our methods use various novel forms of Markov models whose states correspond to users and queries in addition to Web pages and whose links also reflect the relationships derived from query-result clicks, query refinements, and explicit ratings.

Keywords: Query logs, link analysis, Markov reward model

Joint work of: Luxenburger, Julia; Weikum, Gerhard

Full Paper: http://drops.dagstuhl.de/opus/volltexte/2007/1066

Full Paper:
http://db.ucsd.edu/webdb2006/camera-ready/paginated/03-117.pdf

Convergence of aggregation/disaggregation iterative methods

Ivo Marek (Czech Technical University, CZ)

Iterative aggregation/disaggregation methods (IAD) belong to competitive tools for computation the characteristics of Markov chains as shown in some publications devoted to testing and comparing various methods designed to this purpose. The IAD are effective in particular when applied to large ill posed problems. One of the purposes of the proposed talk is to contribute to a possible explanation of this fact. The novelty may consist of the fact that the IAD algorithms do converge independently of whether the iteration matrix of the corresponding process is primitive or not. Some numerical tests are presented and possible applications mentioned.

Keywords: Iterative aggregation/disaggregation method, stochastic matrix

Convergence of iterative aggregation/disaggregation methods based on splittings with cyclic iteration matrices

Ivo Marek (Czech Technical University, CZ)

Iterative aggregation/disaggregation methods (IAD) belong to competitive tools for computation the characteristics of Markov chains as shown in some publications devoted to testing and comparing various methods designed to this purpose.

According to Dayar T., Stewart W.J. *Comparison of partitioning techniques for two-level iterative solvers on large, sparse Markov chains.* SIAM J. Sci. Comput.Vol **21**, No. 5, 1691-1705 (2000), the IAD methods are effective in particular when applied to large ill posed problems. One of the purposes of this paper is to contribute to a possible explanation of this fact. The novelty may consist of the fact that the IAD algorithms do converge independently of whether the iteration matrix of the corresponding process is primitive or not. Some numerical tests are presented and possible applications mentioned; e.g. computing the PageRank.

Keywords: Iterative aggregation methods, stochastic matrix, stationary probability vector, Markov chains, cyclic iteration matrix, Google matrix, PageRank

Joint work of: Marek, Ivo; Pultarová, Ivana; Mayer, Petr

Full Paper: http://drops.dagstuhl.de/opus/volltexte/2007/1067

Parallel Graph Clustering Based on Disturbed Diffusion

Henning Meyerhenke (Universität Paderborn, D)

Graph clustering refers to the placement of nodes into meaningful groups based on some similarity measure. It is an important task in a wide variety of applications, e. g., network analysis for community detection. We propose a new heuristic for this problem, which is a variant of an algorithm that has been successfully applied in shape-optimizing graph partitioning. This iterative approach resembles Lloyd's k-means algorithm. However, its similarity measure is derived from a disturbed diffusion scheme which shows some connections to random walks. This scheme ensures the desired cluster property, namely, that cluster centers are drawn to dense regions of the graph, while the boundaries are in sparse ones.

We present a rigorous proof relying on a potential function that our iterative graph clustering algorithm converges and describe the algorithm's parallel implementation, which uses message-passing communication and includes a multilevel approach to improve convergence speed and solution quality. Experiments show promising results not only for graphs, but also for geometric data sets. Comparisons to related work are drawn on a conceptual level and, where appropriate, on an experimental basis.

Keywords: Disturbed diffusion, graph clustering, parallel data mining

Joint work of: Meyerhenke, Henning; Monien, Burkhard; Sauerwald, Thomas

On algebraic multilevel methods for non-symmetric systems

Reinhard Nabben (TU Berlin, D)

Here we analyze algebraic multilevel methods for solving a linear system of equations whose system matrix is a nonsingular or singular M-matrix.

Such kind of linear systems occur in numerical analysis, probability, economics and operations research. Systems with singular M-matrices arise for example in finite Markov chains where the stationary distribution vector has to be computed. We consider two types of multilevel approximate block factorizations. The first one is known as the AMLI method.

The second method is the multiplicative counterpart of the AMLI method which we call multiplicative algebraic multilevel method, the MAMLI method.

We establish convergence results and comparison results between these methods.

Keywords: multilevel methods, non-symmetric systems, M-matrices

Joint work of: Nabben, Reinhard; Mense, Christian

Extrapolation and minimization procedures for the PageRank vector

Michela Redivo-Zaglia (Universitá di Padova, I)

An important problem in Web search is to determine the importance of each page. This problem consists in computing, by the power method, the left principal eigenvector (the PageRank vector) of a matrix depending on a parameter c which has to be chosen close to 1.

However, when c is close to 1, the problem is ill-conditioned, and the power method converges slowly. So, the idea developed in this paper consists in computing the PageRank vector for several values of c, and then to extrapolate them, by a conveniently chosen rational function, at a point near 1. The choice of this extrapolating function is based on the mathematical considerations about the PageRank vector.

Keywords: Extrapolation, PageRank, Web matrix, eigenvector computation.

Joint work of: Brezinski, Claude; Redivo-Zaglia, Michela

Extended Abstract: http://drops.dagstuhl.de/opus/volltexte/2007/1068

Large scale matrix approximation

Tamas Sarlos (Hungarian Academy of Sciences, H)

In our talk we give efficient approximation algorithms for personalized PageRank and Singular Value Decomposition (SVD).

There are three common properties in these results. First, the algorithms can be applied to extremely large datasets, such as graphs or matrices with millions or billions of nodes. Secondly, it is proven that their approximation error can be made arbitrarily small. Thirdly, efficiency is achieved by novel use of low-space synopsis data structures originally invented for data stream computation.

Personalized PageRank expresses link-based page quality around user selected pages. We achieve unrestricted personalization by combining rounding and randomized sketching techniques in the dynamic programming algorithm of Jeh and Widom. As a theoretical contribution we show that our algorithms use an optimal amount of space by also improving earlier asymptotic worst-case lower bounds. Experimental evaluation, extensions to similarity search and other PageRank like measures expressible as matrix power series are also to be discussed briefly.

Recently several procedures appeared for speeding up SVD computation; most of them are based on random sampling. Our key idea is that low dimensional embeddings can be used to eliminate data dependence inherent in sampling and provide more versatile, linear time pass efficient matrix computation. We present an efficient 2 pass relative error approximation algorithm for singular value decomposition.

Keywords: PageRank, SimRank, SVD, regression, sketches

Google Pageranking Problem: The Model and the Analysis

Stefano Serra Capizzano (Università dell'Insubria di Como, I)

Let A be a given n-by-n complex matrix with eigenvalues $\lambda, \lambda_2, \ldots, \lambda_n$. Suppose there are nonzero vectors $x, y \in \mathbb{C}^n$ such that $Ax = \lambda x$, $y^*A = \lambda y^*$, and $y^*x = 1$. Let $v \in \mathbb{C}^n$ be such that $v^*x = 1$, let $c \in \mathbb{C}$, and assume that $\lambda \neq c\lambda_j$ for each $j = 2, \ldots, n$. Define $A(c) := cA + (1-c)\lambda x v^*$. The eigenvalues of $A(c)$ are $\lambda, c\lambda_2, \ldots, c\lambda_n$. Every left eigenvector of $A(c)$ corresponding to λ is a scalar multiple of $y - z(c)$, in which the vector $z(c)$ is an explicit rational function of c. If a standard form such as the Jordan canonical form or the Schur triangular form is known for A, we show how to obtain the corresponding standard form of $A(c)$.

The web hyper-link matrix $G(c)$ used by Google for computing the PageRank is a special case in which A is real, nonnegative, and row stochastic (taking into consideration the dangling nodes), $c \in (0,1)$, x is the vector of all ones, and v is a positive probability vector. The PageRank vector (the normalized dominant left

eigenvector of $G(c)$) is therefore an explicit rational function of c. Extrapolation procedures on the complex field may give a practical and efficient way to compute the PageRank vector when c is close to 1.

A discussion on the model, on its adherence to reality, and on possible variations is also considered.

Keywords: Google matrix, rank-one perturbation, Jordan canonical form, extrapolation formulae.

Full Paper: http://drops.dagstuhl.de/opus/volltexte/2007/1069

See also: "Jordan canonical form of the Google matrix: a potential contribution to the PageRank computation", SIAM Journal on Matrix Analysis and Applications, Vol. 27, N.2 (2005), pp. 305–312

About graph similarity and applications

Paul Van Dooren (University of Louvain, B)

We present the notion of similarity matrix between two graphs G_A and G_B as the matrix S whose element $s_{i,j}$ measures the similarity of node i in graph G_A and node j in graph G_B. This matrix can be obtained as a solution of the matrix equation $\rho S = ASB' + A'SB$, where A and B are the adjacency matrices of the two graphs.

We then show how to construct low rank approximations of this matrix for large scale graphs. This amounts to an optimization algorithm on isometries, and we describe a fixed point algorithm to solve this optimization problem. We give variants of this problem and describe a number of applications.

Keywords: Large scale graphs, similarity matrix, low rank approximation, optimization

Graph matching with type constraints on nodes and edges

Paul Van Dooren (University of Louvain, B)

In this paper, we consider two particular problems of directed graph matching. The first problem concerns graphs with nodes that have been subdivided into classes of different type. The second problem treats graphs with edges of different types. In the two cases, the matching process is based on a constrained projection of the nodes and of the edges of both graphs in a lower dimensional space. The procedures are formulated as non-convex optimization problems. The objective functions use the adjacency matrices and the constraints on the problem impose the isometry of the so-called projections. Iterative algorithms are proposed to solve the optimization problems. As illustration, we give an example of graph matching for graphs with two types of nodes and graphs with two types of edges.

Keywords: Graph matching, optimization, typed nodes, typed edges

Joint work of: Fraikin, Catherine; Van Dooren, Paul

Full Paper: http://drops.dagstuhl.de/opus/volltexte/2007/1071

PageRank as a function of the damping factor

Sebastiano Vigna (Università di Milano, I)

PageRank is defined as the stationary state of a Markov chain. The chain is obtained by perturbing the transition matrix induced by a web graph with a damping factor α that spreads uniformly part of the rank. The choice of α is eminently empirical, and in most cases the original suggestion $\alpha = 0.85$ by Brin and Page is still used. We give the first mathematical analysis of PageRank when α changes. In particular, we show that, contrarily to popular belief, for real-world graphs values of α close to 1 do not give a more meaningful ranking. We also use resolvent theory to shed some light on the limit values. Then, we give closed-form formulae for PageRank derivatives of any order, and an extension of the Power Method that approximates them with convergence $O(t^k \alpha^t)$ for the k-th derivative. Finally, we show a tight connection between iterated computation and analytical behaviour by proving that the k-th iteration of the Power Method gives exactly the PageRank value obtained using a Maclaurin polynomial of degree k.

Keywords: PageRank, Markov chains

Joint work of: Vigna, Sebastiano; Boldi, Paolo; Santini, Massimo

Full Paper:
http://vigna.dsi.unimi.it/papers.php#BSVPFDF

A Deeper Investigation of PageRank as a Function of the Damping Factor

Sebastiano Vigna (Università di Milano, I)

PageRank is defined as the stationary state of a Markov chain. The chain is obtained by perturbing the transition matrix induced by a web graph with a damping factor α that spreads uniformly part of the rank. The choice of α is eminently empirical, and in most cases the original suggestion $\alpha = 0.85$ by Brin and Page is still used.

In this paper, we give a mathematical analysis of PageRank when α changes. In particular, we show that, contrarily to popular belief, for real-world graphs values of α close to 1 do not give a more meaningful ranking. Then, we give closed-form formulae for PageRank derivatives of any order, and by proving

that the k-th iteration of the Power Method gives exactly the PageRank value obtained using a Maclaurin polynomial of degree k, we show how to obtain an approximation of the derivatives. Finally, we view PageRank as a linear operator acting on the preference vector and show a tight connection between iterated computation and derivation.

Keywords: PageRank, damping factor, Markov chains

Joint work of: Boldi, Paolo; Santini, Massimo; Vigna, Sebastiano

Full Paper: http://drops.dagstuhl.de/opus/volltexte/2007/1072

Stanford Matrix Considered Harmful

Sebastiano Vigna (Università di Milano, I)

I discuss the implications of using small data sets for experiments related to the web graph.

Keywords: Weg graph, PageRank, HITS

Full Paper: http://drops.dagstuhl.de/opus/volltexte/2007/1058

An SVD approach to identifying meta-stable states of Markov chains

Elena Virnik (TU Berlin, D)

Graph clustering algorithms and also spectral graph partitioning already find a broad variety of applications in web information retrieval. Our approach follows the idea of graph clustering and it is tailored to identify meta-stable states of a Markov chain. The presented algorithm reorders the transition matrix of a Markov chain such that the permuted matrix reveals a block structure with high transition probabilities within the blocks and small transition probabilities between different blocks. Originally, our study was motivated by the problem of the identification of meta-stable states of Markov chains in conformation dynamics of molecules. Previous work on this topic usually involves the computation of the eigenvalue cluster close to one, as well as the corresponding eigenvectors and the stationary probability distribution of the stochastic matrix. As a possible less costly alternative, we present an SVD approach to identifying meta-stable states of a stochastic matrix, where we only have to calculate the singular vector corresponding to the second largest singular value. We outline some theoretical background and discuss the advantages of this strategy. Some simulated and real numerical examples that illustrate the effectiveness of the algorithm are presented.

Keywords: Markov chains; singular value decomposition; clustering

Joint work of: Virnik, Elena; Fritzsche, David; Mehrmann, Volker; Szyld, Daniel B.

Term Dependence Issues in Document Retrieval

Hugo Zaragoza (Yahoo Research - Barcelona, E)

Most ranking models make strong assumptions of term independence with respect to relevance; in particular the score of a document is thought to be linear with respect to the score of the different query terms.

These assumptions are reasonable in many traditional IR applications, where documents are considered flat. However, as we increase the complexity of the document representation, term dependence effects grow. This is already the case in Web and corporate search applications, where term dependence effects are clearly hurting search relevance; as we move towards more semantic search applications, term dependence effects only get worse.

In my talk I will discuss the term dependence problem and give an overview of proposed solutions (including our own recent work on this subject), and discuss how this may affect ranking function architecture.