Computing and Informatics, Vol. 31, 2011, 201–204

SPECIAL SECTION ON INFORMATION RETRIEVAL BY MATRIX METHODS ON SUPERCOMPUTER SYSTEMS

Marian Vajteršic

Department of Computer Sciences
University of Salzburg, Austria
&
Department of Informatics, Mathematical Institute
Slovak Academy of Sciences, Bratislava, Slovakia
e-mail: marian@cosy.sbg.ac.at

This special section includes three of the papers submitted and presented at the international workshop Information Retrieval by Matrix Methods on Supercomputer Systems – IRRM '10, which was organized by Marian Vajteršic (University of Salzburg, Austria), Michael W. Berry (University of Tennessee, USA) and Efstratios Gallopoulos (University of Patras, Greece) within the 24$^{\text{th}}$ ACM International Conference on Supercomputing (ICS '10) in Tsukuba, Japan.

The first motivation for the workshop was to present matrix-based methods enabling to quickly and reliably elaborate, evaluate and interpret an increasing amount of data produced by modern society. Especially, extracting a concrete useful information from huge data sets, like web-pages, libraries, image collections is a highly actual topic of computer science and applied mathematics. Since Information Retrieval (IR) over such data is an important application domain for High-Performance Computing (HPC), the other motivation was to discuss the implementation of the matrix methods relevant for solving IR problems on state-of-the-art supercomputer systems and infrastructures like multicores, GPUs and Grids.

The algorithmic foundations of IR depend on linear and multilinear algebra theory and algorithms. For example, systems that rapidly extract information from advanced technology networks like sensors on satellites often make use of low-rank decompositions of large, sparse data structures in matrix or tensor form. These decompositions usually involve the computation of eigenvectors or singular vectors and fast, scalable approximation to such vectors is important for the underlying scheme to be practical for large data sets.

A major challenge is that the scale of IR problems can be enormous and the data fluid. To assure the delivery of accurate responses to queries and other IR opera-

tions within acceptable time constraints, it is critical that the design of algorithms and the engineering of their implementation on HPC platforms combine effective dimensionality reduction, IR specific domain knowledge, and efficient exploitation of the software, architecture and hardware infrastructures.

Some of relevant themes related to IR:

- Innovative Matrix and Tensor-Based Models for IR
- Fast Linear Algebra Solvers and Environments for IR
- Solving Update-Downdate Problem of IR
- Specialized Approaches for Solving Sparse- and Dense- Vector IR Applications
- High-Performance Implementations of IR Matrix-Algorithms
- Solving IR-Problems on a Grid
- Formal Comparisons of Matrix-Based to Matrix-Free Methods.

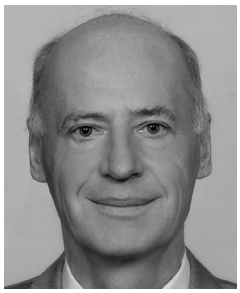
The papers selected for this special section reflect some of these themes in detail. Andreas Janecek, Stefan Schulze Grotthoffand and Wilfried Gansterer (University of Vienna, Austria) discuss in their paper, entitled as "libNMF – A Library for Nonnegative Matrix Factorization" one of the most actual low-rank approximation method and its software realization. This is the Nonnegative Matrix Factorization (NMF) and its key characteristic is that it preserves the nonnegativity also for factors of the low-rank representation. It is a valuable property, because for collected data, which are often nonnegative, it is often required that the low-rank representation is also nonnegative (e.g. in order to avoid some contradictions and/or to make the data interpretation easier). Classical tools, like the Singular Value Decomposition (SVD), cannot guarantee to maintain the nonnegativity. The paper presents a software library called libNMF that provides efficient implementations of several NMF routines. It contains various state-of-the-art NMF algorithms for computing NMF found in the literature, and some methods for initializing the NMF matrix-factors in order to speed-up the convergence. libNMF is written in C and calls external routines from the software libraries LAPack and ARPack , and is based on the BLAS (Basic Linear Algebra Subprograms). The routines of the library are evaluated in terms of computational efficiency and numerical accuracy and compared with the best existing codes available. For using, further developments and testing, the documented source code of libNMF, some test data sets, and detailed documentation of the library are publicly available.

Another view on the reduction of the dimensionality of a large data corpus offers the graph-theory. Concretely, Geoffrey Sanders and Van Emden Henson from Lawrence Livermore National Laboratory, together with Hans De Sterck from University of Waterloo describe in their contribution "Multilevel Aggregation Methods for Small-World Graphs with Application to Random-Walk Ranking" the application of aggregation, which belongs to a graph coarsening techniques. A primary motivation of the authors was to extend the applicability of aggregation to problems on small-world graphs, with the goal of developing these methods for tasks

such as IR. Since such applications can be viewed as large networks consisting of huge number of connected nodes, the coarsening of its graph representation is useful to perform extraction of desired items in a parallel scalable way. On the coarse version of the original graph, multilevel hierarchies can be formed, which allow rapid calculation of low-rank approximations. It is shown that created hierarchies can be used to accelerate convergence properties of processes, related to data retrieval. The paper shows examples of small-world graphs where neighborhood aggregation achieved multilevel hierarchies of optimal complexity. The authors intend to generalize their algorithm to be robust and scalable for a wider class of complicated network graphs. In addition, they are working on related algorithms that compute multiple eigenvectors that are to be used for several types of network calculations relevant to IR retrieval.

The paper "Parallel Retrieval of Dense Vectors in the Vector Space Model", by Tobias Berka and Marian Vajteršic from the University of Salzburg, deals with parallelization of the data retrieval process for the case of dense vectors. Since most text retrieval applications operate on sparse but very high dimensional vectors, most of the existing techniques are based on sparse vector methods. However, there are also text and multimedia retrieval applications which require dimensionality-reduction techniques for data represented by dense and relatively short vectors. As the authors claim, the discrepancy between the traditional focus on sparse vector processing on one hand, and the importance of dense vectors for multimedia retrieval and dimensionality reduction techniques on the other hand has indeed left a gap in parallel and distributed information retrieval methods. Considering the vector space model, they try to fill up this gap with a a design of a parallelization strategy for efficient query processing of data sets formed by dense matrices and vectors. The corpus matrix is a block partitioned across both documents and features and correspondingly distributed on a mesh of processing hosts of a parallel distributed-memory system. The vector-matrix multiplication, which dominates the complexity of retrieval for dense vectors in the vector space model, is performed in parallel across all feature groups. The results are aggregated through a parallel merge-sort across the document groups. The performance of the algorithm was analyzed theoretically and an optimal, linear speed-up in terms of the number of processors has been estimated. The performance for MPI (Message Passing Interface) implementations delivered even super-linear speedup for some splitting schemes.

We hope that these three submissions can illustrate to an interested reader a vital and manifold research activity in the challenging area of automated IR approaches.



Marian Vajteršic graduated in numerical mathematics from Comenius University, Bratislava (Slovak Republic) in 1974. He received his C. Sc. (candidate of sciences) degree in mathematics from the same university in 1984 and he defended the Dr. Sc. (doctor of sciences) degree in there 1997. In 1995, he obtained the habilitation degree in numerical mathematics and parallel processing from the University of Salzburg (Austria). His research activity is focused on the area of parallel numerical algorithms for high-performance computer systems. He is the author of two monographs, co-author of four other books and of more than 100 scientific papers. Since 1974, he is with the Slovak Academy of Sciences in Bratislava, Slovakia. As a Visiting Professor he had been with the universities of Vienna, Bologna, Milan, Linz, Salzburg, Amiens and Munich. Since 2002 he is a Full Professor at the Department of Computer Sciences at the University of Salzburg, Austria.