Final Report on DOE contract #DE-FG03-97ER25325

E.R. Jessup
University of Colorado

This is a report of the results of our work supported by DOE contract #DE-FG03-97ER25325, 7/15/97-7/14/99. We proposed to study three numerical problems. They are 1) the extension of the PMESC parallel programming library, 2) the development of algorithms and software for certain generalized eigenvalue and singular value (SVD) problems, and 3) the application of techniques of linear algebra to an information retrieval technique known as latent semantic indexing (LSI).

The first problem is PMESC–a library for managing task parallel computations on distributed-memory computers. The goal of PMESC is to make it easy for users to write efficient and portable codes for these complex problems on these complex machines. Development of the first version of PMESC is complete (as a thesis quality project [3].) My collaborator Silvia Crivelli is using it and the concepts behind it for problems in computational chemistry under study at Lawrence Berkeley National Laboratory, where she holds a postdoctoral research appointment. During the period of the 1997 DOE award, the following paper about PMESC was accepted for publication:

S. Crivelli and E.R. Jessup, *The PMESC Programming Library for Distributed-Memory MIMD Computers*, J. Parallel and Distributed Computing, 57, 295-321 (1999).

One other result of this work is a new algorithm for parallel termination detection that underlies one of the library routines. In order to be a useful computational tool, the termination detection routine must operate concurrently with the main computation, adding minimal overhead, and it must promptly and correctly detect termination when it occurs [7]. In this paper, we present a new algorithm for detecting the termination of a parallel computation on distributed-memory MIMD computers that satisfies all of those criteria. Our algorithm adapts to system loads, imposes little overhead, and does not depend on any assumptions about the physical interconnection topology of the processors or the specifics of the distributed computation. It is easier to implement and requires only half as many tree traverses as does the most efficient alternative.

We have submitted a paper about our algorithm to *Information Processing Letters*.

11/2 4218 9004 5893 DOE Oakland Oper. 1st Class Fred Howe

DISCLAIMER

The second problem concerns issues in the accurate computation of the SVD and its variants. A blanket result from these papers is that a single, unified approach can be applied to such problems. DOE-supported postdoc Zlatko Drmač and I are presently completing a software template for highly accurate high-performance solution of generalized eigenvalue and singular value problems. Supporting work completed to date includes the following papers:

1. Z. Drmač and E.R. Jessup, *On accurate generalized singular value computation in floating-point arithmetic*, SIAM J. Matrix Anal. Appl., to appear.

2. Z. Drmač, *Accurate computation of the product induced singular value decomposition with applications*, SIAM J. Numer. Anal., to appear.

3. Z. Drmač, *New accurate algorithms for singular value decomposition of matrix triplets*, SIAM J. Matrix Anal. Appl., to appear.

4. Z. Drmač and E.R. Jessup, *On Stable Generalized Singular Value Computation* , in the Proceedings of the the Householder Meeting on Numerical Algebra XIII, 1996, 4 pages.

The third problem is an explanation of the mathematics underlying the method of LSI. LSI is a variant of the *vector space model* [6, 2] of IR in which each document is encoded as a vector, where each vector component reflects the importance of a particular term in representing the semantics of that document. The vectors for all documents in a database are stored as the columns of a single matrix. In LSI, the SVD is used to reduce the rank of that matrix. We proposed to examine a set of algorithmic and software concerns pertaining to the use of LSI in practice. Our study of LSI, however, took us in an unexpected direction. We tried to begin our work with a tutorial paper on LSI, designed for use in the classroom, but we found that the equations underlying LSI as defined in [1, 4] had no reasonable geometric interpretation. They had been designed for the express purpose of allowing direct comparison between the documents and the terms used to describe them, despite the fact that terms and documents are typically represented by vectors of different dimension. That realization led us to develop a new method of IR that applies the vector space model directly to a low-rank approximation of the database matrix. The operations performed in this version of the vector space model admit an easier geometric interpretation than do those underlying LSI. That method is described in the following paper:

M. Berry, Z. Drmač, and E.R. Jessup, *Using Linear Algebra for Information Retrieval*, SIAM Review, 41, 335-362 (1999).

We are presently beginning a study of how the method performs in practice.

A Ph.D. thesis on peripheral issues in LSI was completed by grad student Laura Mather [5]. She was a full time employee of the National Security Agency while still a student, and the topic of her thesis was determined by the NSA. Because her research evolved from a mathematical topic to one in information retrieval, Laura's advisors during the last year of her thesis work were Jim Martin of the University of Colorado and Joe Morgan of the NSA.

References

[1] M. Berry, S. Dumais, and G. O'Brien, *Using linear algebra for intelligent information retrieval*, SIAM Review, 37 (1995), pp. 573–595.

[2] C. Buckley, G. Salton, J. Allan, and A. Singhanl, *Automatic query expansion using SMART: TREC 3*, in Overview of the Third Text REtrieval Conference, D. Harman, ed., National Institute of Standards and Technology Special Publication 500-226, April 1995.

[3] S. Crivelli, *A programming paradigm for distributed-memory computers*, PhD thesis, Dept. of Computer Science, University of Colorado at Boulder, 1995.

[4] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman, *Indexing by latent semantic analysis*, Journal of the American Society for Information Science, 41 (1990), pp. 391–407.

[5] L. Mather, *Enhancing Cluster-Based Retrieval through Linear Algebra*, PhD thesis, University of Colorado, 1998.

[6] G. Salton and M. McGill, *Introduction to Modern Information Retrieval*, McGraw Hill, New York, 1983.

[7] A. Sinha, L. Kale, and B. Ramkumar, *A dynamic and adaptive quiescence detection algorithm*, Tech. Rep. 93-11, Department of Computer Science, University of Illinois, 1993.