SANDIA REPORT

SAND2010-6357
Unlimited Release
Printed September 2010

LDRD Final Report: Leveraging Multi-way Linkages on Heterogeneous Data

Daniel M. Dunlavy and Tamara G. Kolda

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Sandia National Laboratories

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

Printed in the United States of America. This report has been reproduced directly from the best available copy.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone:	(865) 576-8401
Facsimile:	(865) 576-5728
E-Mail:	reports@adonis.osti.gov
Online ordering:	http://www.osti.gov/bridge

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd
Springfield, VA 22161

Telephone:	(800) 553-6847
Facsimile:	(703) 605-6900
E-Mail:	orders@ntis.fedworld.gov
Online ordering:	http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online

SAND2010-6357
Unlimited Release
Printed September 2010

LDRD Final Report: Leveraging Multi-way Linkages on Heterogeneous Data

Tamara G. Kolda
Informatics and Systems Assessments Department
Sandia National Laboratories
Livermore, CA 94551-9159
Email: tgkolda@sandia.gov

Daniel M. Dunlavy
Computer Science & Informatics Department
Sandia National Laboratories
Albuquerque, NM 87123-1318
Email: dmdunla@sandia.gov

Abstract

This report is a summary of the accomplishments of the "Leveraging Multi-way Linkages on Heterogeneous Data" which ran from FY08 through FY10. The goal was to investigate scalable and robust methods for multi-way data analysis. We developed a new optimization-based method called CPOPT for fitting a particular type of tensor factorization to data; CPOPT was compared against existing methods and found to be more accurate than any faster method and faster than any equally accurate method. We extended this method to computing tensor factorizations for problems with incomplete data; our results show that you can recover scientifically meaningfully factorizations with large amounts of missing data (50% or more). The project has involved 5 members of the technical staff, 2 postdocs, and 1 summer intern. It has resulted in a total of 13 publications, 2 software releases, and over 30 presentations. Several follow-on projects have already begun, with more potential projects in development.

Acknowledgments

This work was fully supported by Sandia's Laboratory Directed Research & Development (LDRD) program.

Contents

This page intentionally left blank.

1 Introduction

The focus of the "Leveraging Multi-way Linkages on Heterogeneous Data" LDRD was investigating and developing novel data analysis methods for heterogeneous data. The motivation is the ubiquity of such data in the intelligence community, though applications also exist in social network, bibliometric, critical infrastructure, and complex biological systems analysis.

Our goal was to investigate techniques for combining heterogeneous entities and the multiple linkages (i.e., relationships) between them. We developed a new class of optimization-based algorithms, emphasizing scalability and robustness, that can be extended to multi-way linkage dimensionality reduction. Such techniques can be used to map heterogeneous entities into a shared conceptual space, which is in turn fundamental for solving a variety of data analysis problems; we have specifically studied an example involving link prediction.

Sandia's expertise in data and graph analysis, matrix and tensor methods, and high-performance computing makes it natural for us to pursue this line of inquiry.

The remainder of this report is organized as follows. Section 2 reviews highlights of our technical contributions. Section 3 summarizes the impact of the LDRD in terms of staff participation, communication of results, and community involvement. Finally, Section 3.6 discusses follow-on projects.

This page intentionally left blank.

2 Technical Impact

Summaries of key results from this LDRD are summarized below. For a complete list of publications, see Section 3.2.

2.1 CPOPT: Optimization for fitting the CP Tensor Decomposition [1]

Tensor decompositions are higher-order analogues of matrix decompositions and have proven to be powerful tools for data analysis. In particular, we are interested in the canonical tensor decomposition, otherwise known as CANDECOMP/PARAFAC (CP), which expresses a tensor as the sum of component rank-one tensors and is used in a multitude of applications such as chemometrics, signal processing, neuroscience, and web analysis. The task of computing CP, however, can be difficult. The typical approach is based on alternating least squares (ALS) optimization, but it is not accurate in the case of overfactoring. High accuracy can be obtained by using nonlinear least squares (NLS) methods; the disadvantage is that NLS methods are much slower than ALS. In this paper, we propose the use of gradient-based optimization methods. We discuss the mathematical calculation of the derivatives and show that they can be computed efficiently, at the same cost as one iteration of ALS. Computational experiments demonstrate that the gradient-based optimization methods are more accurate than ALS and faster than NLS in terms of total computation time.

2.2 CPWOPT: Fitting the CP Tensor Decomposition with Incomplete Data [3]

The problem of incomplete data—i.e., data with missing or unknown values—in multi-way arrays is ubiquitous in biomedical signal processing, network traffic analysis, bibliometrics, social network analysis, chemometrics, computer vision, communication networks, etc. We consider the problem of how to factorize data sets with missing values with the goal of capturing the underlying latent structure of the data and possibly reconstructing missing values (i.e., tensor completion). We focus on one of the most well-known tensor factorizations that captures multi-linear structure, CANDECOMP/PARAFAC (CP). In the presence of missing data, CP can be formulated as a weighted least squares problem that models only the known entries. We develop an algorithm called CP-WOPT (CP Weighted OPTimization) that uses a first-order optimization approach to solve the weighted least squares problem. Based on extensive numerical experiments, our algorithm is shown to successfully factorize tensors with noise and up to 99% missing data. A unique aspect of our approach is that it scales to sparse large-scale data, e.g., $1000 \times 1000 \times 1000$ with five million known entries (0.5% dense). We further demonstrate the usefulness of CP-WOPT on two real-world applications: a novel EEG (electroencephalogram) application where missing data is frequently encountered due to disconnections of electrodes and the problem of modeling computer network traffic where data may be absent due to the expense of the data collection process.

2.3 Temporal Link Prediction [10]

The data in many disciplines such as social networks, web analysis, etc. is link-based, and the link structure can be exploited for many different data mining tasks. In this paper, we consider the problem of temporal link prediction: Given link data for times 1 through T, can we predict the links at time $T+1$? If our data has underlying periodic structure, can we predict out even further in time, i.e., links at time $T+2$, $T+3$, etc.? In this paper, we consider bipartite graphs that evolve over time and consider matrix- and tensor-based methods for predicting future links. We present a weight-based method for collapsing multi-year data into a single matrix. We show how the well-known Katz method for link prediction can be extended to bipartite graphs and, moreover, approximated in a scalable way using a truncated singular value decomposition. Using a CANDECOMP/PARAFAC tensor decomposition of the data, we illustrate the usefulness of exploiting the natural three-dimensional structure of temporal link data. Through several numerical experiments, we demonstrate that both matrix- and tensor-based techniques are effective for temporal link prediction despite the inherent difficulty of the problem. Additionally, we show that tensor-based techniques are particularly effective for temporal data with varying periodic patterns.

2.4 Fast Tensor Computations in Parallel and on GPUs [7]

The tensor eigenproblem has many important applications, and both mathematical and application-specific communities have taken recent interest in the properties of tensor eigenpairs as well as methods for computing them. In particular, Kolda and Mayo [13] present a generalization of the matrix power method for symmetric tensors. We focus in this work on efficient implementation of their algorithm, known as the shifted symmetric higher-order power method, and on how a GPU can be used to accelerate the computation up to 70× over a sequential implementation for an application involving many small tensor eigenproblems.

3 Impact

3.1 Staff Participation

The primary technical staff involved in this project were the authors of this report; however, other staff, postdocs, and interns were involved as listed below.

- Tamara Kolda (8966) — Principal investigator, FY08–FY10
- Daniel Dunlavy (1415) — Primary contributor to all aspects of the project, FY08–FY10
- Ann Yoshimura (8116) — Database support, FY08
- Evrim Acar (8962, postdoc) — Full-time postdoc on this project, FY09–FY10
- Todd Plantenga (8958) — Optimization consultant, FY10
- Nicole Lemaster (8961) — Parallel implementation of tensor decomposition, FY10
- David Gleich (8966, Von Neumann postdoc) — Entity resolution and disambiguation, FY10
- Grey Ballard (8966, summer intern) — GPU implementation of tensor algorithms, Summer FY10

3.2 Publications

This LDRD produced 13 publications, broken down below.

- 7 journal articles
  - ACM Transactions on Knowledge Discovery [10]
  - Chemometrics and Intelligent Laboratory Systems [3]
  - IEEE Transactions on Knowledge and Data Engineering [5]
  - Journal of Chemometrics [8, 1]
  - SIAM Journal on Scientific Computing [6]
  - SIAM Review [12]
- 3 refereed conference and workshop articles
  - ICDM'08: IEEE International Conference on Data Mining [14]
  - LDMTA'09: 1st Workshop on Large-Scale Data Mining: Theory and Applications [2]
  - SDM'10: SIAM International Conference on Data Mining [4]
- 1 book chapter [9]
- 2 technical reports [11, 7]

3.3 Software

This LDRD produced two software releases.

- Tensor Toolbox for MATLAB, Version 2.4 (released 2010)
- Poblano for MATLAB, Version 1.0 (released 2010)

3.4 Presentations

This LDRD produced over 30 presentations, with highlights listed below.

- Keynote Invited Talks at Major Conferences
  - IEEE International Conference on Data Mining (ICDM'07), Omaha, Nebraska, Oct. 28–31, 2007
  - BIT 50 Conference - Trends in Numerical Computing, Lund, Sweden, June 17-20, 2010
  - BIRS workshop on Sparse Random Structure, Banff, Canada, Jan. 24–29, 2010
- Keynote Invited Talks at Minor Conferences/Workshops

- Symposium on Gene Golub's Legacy: Matrix Computations — Foundation and Future, Stanford University, California, Mar. 1, 2008
  - ICDM09 Workshop on Large Scale Data Mining: Theory and Applications (LDMTA), Miami Beach, Florida, Dec. 6, 2009
  - Pete Stewart
- Invited Talks at International Workshops
  - GAMM Seminar on Tensor Approximations, Max-Planck Institute for Mathematics in the Sciences, Leipzig, Germany, Jan. 25–26, 2008
  - Computational Algebraic Statistics, Theories and Applications (CASTA2008), Kyoto, Japan, Dec. 10–11, 2008
- Invited Talks at U.S. Workshops
  - Numerical Tools and Fast Algorithms for Massive Data Mining, Search Engines and Applications, IPAM, UCLA, Los Angeles, CA, Oct. 22–26, 2007
  - Multi-Manifold Data Modeling and Applications, Institute for Mathematics and Its Applications (IMA), Minneapolis, Minnesota, Oct. 27–30, 2008
  - Future Directions in Tensor-Based Computation and Modeling, NSF, Arlington, Virginia, Feb. 20–21, 2009 (two talks)
  - BIRS workshop on Sparse Random Structure, Banff, Canada, Jan. 24–29, 2010

3.5 Community Involvement

The staff from this project were involved with numerous events, chairing meetings, serving on program committees, and organizing minisympoia. Of particular note, we co-organized the AIM Workshop on Computational Optimization for Tensor Factorizations in Palo Alto, California in 2010.

3.6 Follow-on Projects

There are two externally-funded (DOE Office of Science Applied Math Program and a WFO project) and one internally-funded project (HSD LDRD) that have already come out of this project. Several more proposals are being developed.

References

[1] E. Acar, D. M. Dunlavy, and T. G. Kolda, *A scalable optimization approach for fitting canonical tensor decompositions*, Journal of Chemometrics. in press.

[2] ——, *Link prediction on evolving data using matrix and tensor factorizations*, in LDMTA'09: Proceeding of the ICDM'09 Workshop on Large Scale Data Mining Theory and Applications, IEEE Computer Society Press, Dec. 2009, pp. 262–269.

[3] E. Acar, D. M. Dunlavy, T. G. Kolda, and M. Mørup, *Scalable tensor factorizations for incomplete data*, Chemometrics and Intelligent Laboratory Systems. in press.

[4] ——, *Scalable tensor factorizations with missing data*, in SDM10: Proceedings of the 2010 SIAM International Conference on Data Mining, Philadelphia, Apr. 2010, SIAM, pp. 701–712.

[5] E. Acar and B. Yener, *Unsupervised multiway data analysis: A literature survey*, IEEE Transactions on Knowledge and Data Engineering, 21 (2009), pp. 6–20.

[6] B. W. Bader and T. G. Kolda, *Efficient MATLAB computations with sparse and factored tensors*, SIAM Journal on Scientific Computing, 30 (2007), pp. 205–231.

[7] G. Ballard, T. G. Kolda, and T. Plantenga, *Efficiently computing tensor eigenvalues on a GPU*, tech. report, Sandia National Laboratories, Albuquerque, NM and Livermore, CA, 2010.

[8] R. Bro, E. Acar, and T. G. Kolda, *Resolving the sign ambiguity in the singular value decomposition*, Journal of Chemometrics, 22 (2008), pp. 135–140.

[9] D. M. Dunlavy, T. G. Kolda, , and W. P. Kegelmeyer, *Multilinear algebra for analyzing data with multiple linkages*, in Graph Algorithms in the Language of Linear Algebra, J. Kepner and J. Gilbert, eds., Fundamentals of Algorithms, SIAM, Philadelphia. in press.

[10] D. M. Dunlavy, T. G. Kolda, and E. Acar, *Temporal link prediction using matrix and tensor factorizations*, ACM Transactions on Knowledge Discovery from Data. in press.

[11] ——, *Poblano v1.0: A matlab toolbox for gradient-based optimization*, Tech. Report SAND2010-1422, Sandia National Laboratories, Albuquerque, NM and Livermore, CA, 2010.

[12] T. G. Kolda and B. W. Bader, *Tensor decompositions and applications*, SIAM Review, 51 (2009), pp. 455–500.

[13] T. G. Kolda and J. R. Mayo, *Shifted power method for computing tensor eigenpairs.* arXiv:1007.1267v1 [math.NA], July 2010.

[14] T. G. Kolda and J. Sun, *Scalable tensor decompositions for multi-aspect data mining*, in ICDM 2008: Proceedings of the 8th IEEE International Conference on Data Mining, Dec. 2008, pp. 363–372.

DISTRIBUTION:

1	MS 0899	Technical Library, 9536 (electronic copy)
1	MS 0123	D. Chavez, LDRD Office, 1011

Sandia National Laboratories