Tensor Decompositions Provide New Methods for Analyzing Critical Data Mining Applications

Contact: Brett W. Bader, Computer Science and Informatics, *bwbader@sandia.gov*

Tensor decomposition techniques have been applied at Sandia to a wide range of problems related to national security, including social network analysis, email surveillance, cross-language information retrieval, and web link analysis.

For many years, data mining has benefited from linear algebraic techniques and matrix decompositions, such as the singular value decomposition (SVD). These techniques have proven successful in challenging problems, from web search to social network analysis. Recently, research at Sandia with funding from the Lab Directed Research and Development (LDRD) program has taken data mining to another dimension, quite literally. Sandia researchers have moved beyond matrices and have formulated data mining problems using multidimensional arrays, which are analyzed subsequently with tensor decomposition [1]. By incorporating metadata in the problem as an extra dimension, these novel techniques extract latent information and subtle relationships in the multivariate data that are missed by matrix techniques. Tensor decompositions techniques have been applied at Sandia to a wide range of problems related to national security, including social network analysis, email surveillance, cross-language information retrieval, and web link analysis.

In joint research with Prof. Michael Berry, of the University of Tennessee, a new approach was developed for automatic conversation detection in email over time where a term-by-author-by-time array encodes an email collection [2]. Tensor analysis of over fifty thousand email messages released in the Enron investigation identified dominant conversations based on the theoretical information gain of terms used in the body of the emails and their co-occurrence information over time and by authors (Figure 1). Strong correlations among terms, authors, and time were identified automatically and, when taken together, represented a conversation. These conversations helped to illuminate activities that occurred just prior to the downfall of Enron.

Another application of tensor decompositions was a data mining validation demonstration that showed how one or two analysts could extract preprogrammed terrorist activities from a large synthetic data set [3]. In this study, we developed a strategy based on algebraic models to extract and sequence important activities and specific events from data sources such as news articles and blogs. The ability to automatically reconstruct a plot or confirm involvement in a questionable activity is greatly facilitated by our approach. We applied our algorithms to the terrorism-based scenarios of the VAST 2007 Contest data set to demonstrate how term-by-entity associations can be used for scenario/plot discovery and evaluation.

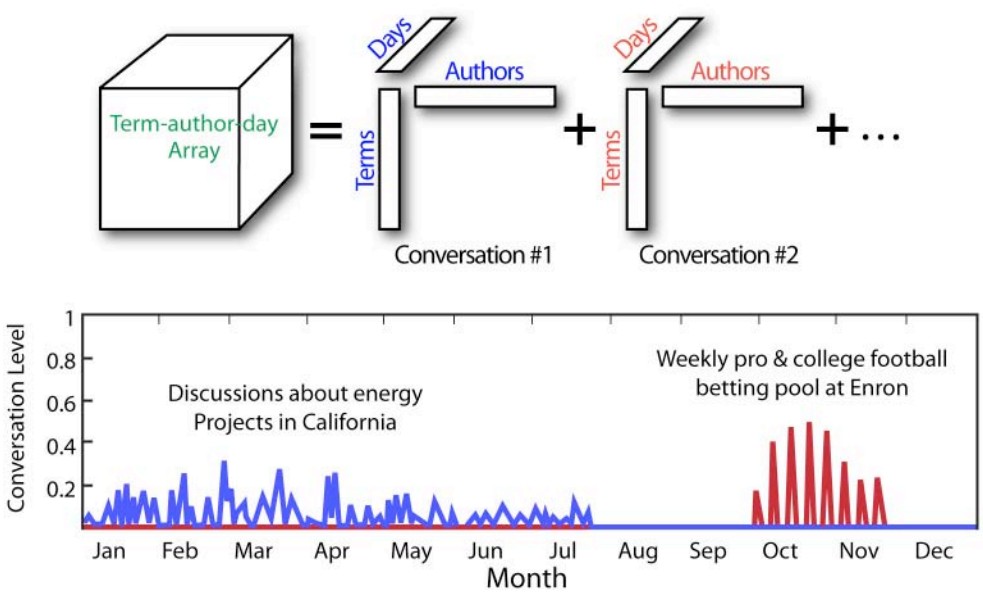

Figure 1. Tensor analysis of Enron emails identified conversations (topic, participants, and temporal profile), ranging from routine business topics to a weekly betting pool involving two dozen employees.

For this study, we set out to automatically detect associations of entities (specific people, locations, organizations, etc.) with terms that appear within the same document and, concurrently, have the same co-occurrence pattern in time. We used the nonnegative tensor factorization (NTF), which is based on the well-known PARAFAC model for multidimensional data, to extract useful term-entity-time associations. These associations provide context for intelligence analysts and provide concise summaries for further investigation. Using the VAST 2007 Contest data set, we have successfully applied our analysis tools to a large, synthetic data set with a "buried" illegal activity and terror scenario.

Our algorithms were implemented in the Tensor Toolbox for MATLAB (*http://csmr.ca.sandia.gov/~tgkolda/TensorToolbox/*), which was created in a previous LDRD project to facilitate research with tensors and now has more than one thousand registered users worldwide. An ongoing effort is to achieve scalable solutions for larger data mining problems using Sandia's expertise in high-performance computing. With current work in our Networks Grand Challenge LDRD, we are developing a tensor decomposition library in C++ for fast algorithms on serial and parallel computers. This advance will enable us to analyze even larger problems, whether from text corpora for content analysis or from network traffic in cyber security. In addition, we are developing different types of tensor decompositions for pattern analysis and clustering in, for example, directed graphs.

References

1. Kolda, T. G., and Bader, B. W. (2009). Tensor Decompositions and Applications, to appear in *SIAM Review*.

2. Bader, B. W., Berry, M., and Browne, M. (2008). Discussion Tracking in Enron email using PARAFAC Part III, Email Surveillance and Filtering, p. 147 in *Survey of Text Mining II*, Springer.

3. Bader, B. W., Puretskiy, A. A., and Berry, M. W. (2008). Scenario Discovery Using Nonnegative Tensor Factorization, *Lecture Notes in Computer Science*, Vol. **5197**, pp. 791-805.

This work has been supported by SNL's Laboratory Directed Research and Development (LDRD) Program.