Mining Large Graphs And Streams Using Matrix And Tensor Tools*

Christos Faloutsos
School of Computer Science
Carnegie Mellon University
christos@cs.cmu.edu

Tamara G. Kolda
Sandia Nat. Labs
tgkolda@sandia.gov

Jimeng Sun
School of Computer Science
Carnegie Mellon University
jimeng@cs.cmu.edu

ABSTRACT

Coevolving streams of numerical measurements, as well as time evolving graphs, can well be represented as tensors. Here we review the fundamental matrix and tensors tools for the analysis and mining of large scale streams and graphs.

Categories and Subject Descriptors

H.2 [**Database Management**]

General Terms

Algorithms, Performance

Keywords

Data Mining, Tensors, Streams

1. DESCRIPTION AND OBJECTIVES

How can we find patterns in sensor streams (eg., a sequence of temperatures, water-pollutant measurements, or machine room measurements)? How can we mine Internet traffic graph over time? Further, how can we make the process incremental? We review the state of the art in four related fields: (a) numerical analysis and linear algebra (b) multi-linear/tensor analysis (c) graph mining and (d) stream mining. We will present both theoretical results and algorithms as well as case studies on several real applications. Our emphasis is on the intuition behind each method, and on guidelines for the practitioner.

2. CONTENT AND OUTLINE

Part I - Fundamentals

- *Data model - Fundamental concepts:* Time series; Matrices Tensors.
- *Matrix analysis:* SVD, PCA and eigen-decomposition; Page-rank, HITS; sparse decompositions and CUR; Co-clustering.
- *Tensor analysis:* Parafac; Tucker Model (Tucker 1 and PCA; Tucker 2 and Tensor PCA; Tucker 3 and High-order SVD); Other models (Combination of PARAFAC and Tucker; DEDICOM).

Part II - Applications

- *Software:* Issues: Scalability, Accuracy, Sparsity.
- *Case studies:* Sensor networks; Machine monitoring; Internet forensic computing; Social network analysis; Web graph study.

3. INTENDED AUDIENCE

Researchers who want to get up to speed with the major tools in stream mining, graph mining. Also, practitioners who want a concise, intuitive overview of the state of the art.

4. ABOUT THE INSTRUCTORS

Christos Faloutsos is a Professor at Carnegie Mellon University. He has received the Research Contributions Award in ICDM 2006, nine "best paper" awards and several teaching awards. His research interests include data mining for streams and networks, fractals, indexing for multimedia and bio-informatics data bases, and performance.
Tamara G. Kolda is a researcher at Sandia National Laboratories in Livermore, California and has received the Presidential Early Career Award for Scientists and Engineers (2003). She has published over 25 refereed articles and released several software packages including the MATLAB Tensor Toolbox. She is an associate editor for the SIAM Journal on Scientific Computing. Her research interests include multilinear algebra and tensor decompositions, data mining, optimization, nonlinear solvers, graph algorithms, parallel computing and the design of scientific software.
Jimeng Sun is a PhD candidate in Computer Science Department at Carnegie Mellon University. His rearch interests include data mining on streams, graphs and tensors, anomaly detection.

5. ACKNOWLEDGMENTS

This work was partially supported by NSF Grants IIS-0326322, IIS-0534205, by PITA (Pennsylvania Infrastructure Technology Alliance), and by DOE (LLNL) contract No.W-7405-ENG-48.

SIGMOD'07, June 12–14, 2007, Beijing, China.
ACM 978-1-59593-686-8/07/0006.