

TEXT MINING 2006

**Proceedings of the Fourth Workshop on Text Mining
Sixth SIAM International Conference on Data Mining
Hyatt Regency Bethesda
Bethesda, Maryland
April 22, 2006
URL: <http://www.siam.org/meetings/sdm06>**

Organizers:

Michael W. Berry (Tennessee) and Malu Castellanos (Hewlett-Packard)



Superior software that gives
you *The Power to Know.*

**SIAM generated table of contents
goes here (page 1)**

OVERVIEW

The proliferation of digital computing devices and their use in communication continues to result in an increased demand for systems and algorithms capable of mining textual data. Thus, the development of techniques for mining unstructured, semi-structured, and fully structured textual data has become quite important in both academia and industry. As a result, a one-day workshop on text mining was held on April 22, 2006 in conjunction with the SIAM Sixth International Conference on Data Mining to bring together researchers from a variety of disciplines to present their current approaches and results in text mining. The workshop surveyed the emerging field of text mining - the application of techniques of machine learning in conjunction with natural language processing, information extraction and algebraic/mathematical approaches to computational information retrieval. Many issues are being addressed in this field ranging from the development of new document classification and clustering models to novel approaches for topic detection and tracking. The goal of this workshop was to provide a venue for researchers to share initial approaches and preliminary results of recent research in text mining. Fifty-four authors representing industry, academia and national research laboratories from 13 different countries submitted a total of 25 papers. After careful review, ten papers and four posters were selected for publication and presentation. Following the success of the previous three editions of the SIAM Text Mining Workshops (TM 2001, TM 2002, TM 2003) this Fourth Edition intends to generate interest and provide insight into the state of the art of text mining. [Following the success of the previous three editions of the SIAM Text Mining Workshop \(TM2001, TM 2002, TM2003\) this](#)

Michael W. Berry
Department of Computer Science
University of Tennessee, Knoxville

Malu Castellanos
Hewlett-Packard Laboratories
Hewlett-Packard, Palo-Alto, CA

ACKNOWLEDGEMENTS

Special thanks to Murray Browne at the University of Tennessee, Knoxville for his assistance in preparing this volume, and to the members of the Program Committee for their diligent efforts in reviewing the 25 manuscripts submitted. The workshop cover image was designed by Jeff Romaniuk at the University of Tennessee, Knoxville. We also appreciate the support of our sponsors PureDiscovery of Dallas, Texas and SAS of Cary, North Carolina.

About SAS

SAS is the market leader in providing a new generation of business intelligence software and services that create true enterprise intelligence. SAS solutions are used at 40,000 sites - including 96 of the top 100 companies on the Fortune Global 500® - to develop more profitable relationships with customers and suppliers; to enable better, more accurate and informed decisions; and to drive organizations forward. SAS is the only vendor that completely integrates leading data warehousing, analytics and traditional BI applications to create intelligence from massive amounts of data. For nearly three decades, SAS has been giving customers around the world The Power to Know®.

About PureDiscovery

PureDiscovery Corporation, based in Dallas, Texas, is a privately held software company. PureDiscovery is the creator of EXgrid, the intelligent grid architecture that transforms existing data repositories into dynamic knowledge and innovation networks. EXgrid creates universal access to virtually any information without disrupting or replacing the clients existing network infrastructure. Organizations placing an emphasis on research, intellectual property, intelligence gathering, collaboration and knowledge sharing can significantly benefit from the use of EXgrid.

Program Committee

Michael W. Berry, University of Tennessee
Malu Castellanos, Hewlett-Packard
Chris Ding, LBNL (NERSC)
William Ferng, Boeing
Steven Soderland, University of Washington
Kyle Gullivan, Florida State University
Peg Howland, Utah State University
Lakshminarayanan Choudur, Hewlett-Packard

Rosie Jones, Yahoo Research Labs
Mei Kobayashi, IBM Tokyo Research Lab
Haesun Park, Georgia Tech
April Kontostathis, Ursinus College
Padma Raghavan, Pennsylvania State University
Efstratios Gallopoulos, University. of Patras
Pierre Senellart, INRIA
Murray Browne, University of Tennessee

Workshop Schedule

Introduction

8:15- 8:20am *Michael W. Berry, Malu Castellanos*

Keynote Speaker

8:20- 9:00am The Needle in the Haystack Problem: Discovering Anomalies
in Text Documents
Dr. Ashok Srivastava, NASA Ames Research Center

Abstract: An important problem that faces many governmental and industrial organizations is that of discovering the description of a recurring phenomenon in text documents. In many applications, the recurring phenomenon has a low frequency of occurrence, thus complicating its discovery. We call such low-frequency events that tend to co-occur “recurring anomalies.” Conventional text mining methods tend to overlook these low-frequency events. The problem of discovering recurring anomalies arises in numerous application domains including fraud, counter-terrorism and security, analysis of complex systems, and warranty and maintenance reports. This talk describes the problem in some detail from a mathematical perspective and then discusses the past and current work in the field. We compare the performance of several existing methods and novel text mining methods that we have developed on text reports regarding complex aerospace systems.

Session I: Topic Detection and Tracking

9:00- 9:30am Confirming Protein-Protein Interactions by Text Mining
David Otasek, Kevin Brown, and Igor Jurisica

9:30-10:00am Entropy Based Measure Functions for Analyzing Time Stamped Documents
Parvathi Chundi, Rui Zhang, and Malu Castellanos

10:00-10:30am Break

Session II: Text Classification

10:30-11:00am One-Sided Non-Negative Matrix Factorization and Non-Negative Centroid
Dimension Reduction for Text Classification
Haesun Park and Hyunsoo Kim

11:00-11:30am Document Author Classification using Generalized Discriminant Analysis
Todd Moon, Peg Howland, and Jacob Gunther

Session III: Poster Presentations

Posters will be on display throughout the workshop and in this session presenters will briefly summarize their work.

11:30-12:15pm Automatically Adjusting Content Taxonomies for Hierarchical Classification
Jianping Zhang, Lei Tang, and Huan Liu

Using Query History to Prune Query Results
Daniel Waegel and April Kontostathis

OPTICS on Text Data: Experiments and Test Results
Shourya Roy and Deepak P

ZIP and Data Document Visualization
Dora Alvarez-Medina and Hugo Hidalgo-Silva

12:15-1:45pm Lunch

Session IV: Clustering Algorithms

1:45-2:15pm Ontology-based Distance Measure for Text Clustering
Liping Jing, Lixin Zhou, Michael K. Ng and Joshua Zhexue Huang

2:15-2:45pm Model-based Overlapping Co-Clustering
Mahdi Shafie and Evangelos Milios

2:45-3:15pm Scaling Clustering Algorithms with Bregman Distances
Jacob Kogan and Marc Teboulle

3:15-3:45pm Break

Session V: LSA and Vector Space Models

3:45-4:15pm Latent Semantic Analysis and Fiedler Embeddings
Bruce Hendrickson

4:15-4:45pm Exploring Term-Document Matrices from Matrix Models in Text Mining
Efstratios Gallopoulos and Ioannis Antonellis,