

## Workshop Schedule

Presenters are listed in *italic*.

### Introduction

8:00-8:15am

*Michael W. Berry*, University of Tennessee, Knoxville

### Keynote

8:15-9:00am

“Using Text Mining to Measure Similarity Between Words and Objects”

*Mehran Sahami*, Google

**Abstract:** The World Wide Web provides a wealth of data that can be harnessed to help improve information retrieval and increase understanding of the relationships between different entities. In many cases, we are often interested in determining how similar two entities may be to each other, where the entities may be pieces of text or descriptions of some object. In this work, we examine multiple instances of this problem, and show how they can be addressed by harnessing data mining techniques applied to large web-based data sets. Specifically, we examine the problems of determining the similarity of short texts (even those that may not share any terms in common) and also of learning similarity functions for semi-structured data to address tasks such as record linkage between objects. While we present rather different techniques for each problem, we show how measuring similarity between entities in these domains has a direct application to the overarching goal of improving information access for users of web-based systems.

### Session I: Factor Analysis

9:00-9:30am

"Discussion Tracking in Enron Email Using PARAFAC"  
*Brett W. Bader, Michael W. Berry, Murray Browne*

9:30-10:00am

"Exploiting Factor Analysis Approximations in Dimension Reduction"  
*Peg Howland*

10:00-10:30am

Break

### Session II: Clustering Algorithms

10:30-11:00am

"K-means Steering of Spectral Divisive Clustering Algorithms"  
*Dimitrios Zeimpekis, Efstratios Gallopoulos*

11:00-11:30am

"Hybrid Clustering of Large High Dimensional Data"  
*Jacob Kogan, Charles Nicholas, Mike Wiacek*

11:30-12:00pm

"Local Semantic Kernels for Clustering of Text Documents"  
*Loulwah AlSumait, Carlotta Domeniconi*

12:00-1:15pm Lunch

### Session III: Contest Winners

1:15-1:30pm Contest Introduction  
*Ashok Srivastava, NASA Ames*

**Contest Procedure:** The competition was organized and judged by members of the Intelligent Data Understanding group at NASA Ames Research Center. A training data set was provided over a month in advance of the deadline, giving the contestants time to develop their approaches. Two days before the deadline the test data set was released. Each contestant submitted their labeling of the test data set, their confidences in the labeling, and source code implementing their approach. The scores of the submissions were calculated using a small Java program that implemented the score function detailed in the contest rules. The program and its source code were released to the contestants prior to the submission deadline so that they could both validate its correctness and use it to tune their algorithms. In addition to scoring the submissions, each contestant's code was run to ensure that it worked and produced the same output that was submitted. This was also done to ensure that the contestants properly followed the rules of the contest.

1:30-2:00pm Contest Paper, 3<sup>rd</sup> Place  
"An Analysis of the Effect of Document Representation on the Classification of Noisy Texts"  
*Narjes Sharif-razavian, Mostafa Keikha, Farhad Oroumchian*

2:00-2:30pm Contest Paper, 2<sup>nd</sup> Place  
Anomaly Detection Using Non-negative Matrix Factorization  
Edward Allan, Michael Horvath, Christopher Kopek, Brian Lamb, Thomas Whaples, *Michael W. Berry*

2:30-3:00p.m Contest Paper, 1<sup>st</sup> Place  
Fast and Confident Probabilistic Categorization  
*Cyril Goutte*

3:00-3:30pm Break

### Session IV: Document Filtering and Classification

3:30-4:00 pm "Spam Filtering Based on Latent Semantic Indexing"  
*Wilfried Gansterer, Andreas Janeczek, Robert Neumayer*

4:00-4:30 pm "A Kernel Method for XML Document Representation in Classification"  
*Zhonghang Xia, Guangming Xing, Houduo Qi, Qi Li*