

Workshop Schedule

Presenters are listed in *italic*.

Introduction

8:00-8:15am

Michael W. Berry, University of Tennessee, Knoxville

Keynote

8:15-9:00am

“Algebraic Techniques for Multilingual Document Clustering”

Brett W. Bader, Sandia National Laboratories

Abstract:

Text documents in multiple languages pose a problem if one wants to cluster them by topic. One approach of translating everything to a common language is not feasible when dealing with a large corpus or many languages. This presentation will show a variety of novel algebraic methods for efficiently clustering multilingual text documents. The methods use a multilingual parallel corpus as a 'Rosetta Stone' from which algorithmic variations of Latent Semantic Analysis (LSA) are able to learn concepts in a multilingual term space. New documents are projected into this concept space to produce language-independent feature vectors for subsequent use in similarity calculations or machine learning applications. Our numerical experiments show that the new methods have better performance than LSA and can be used in machine learning tasks.

Session I:

Text Streams

9:00-9:30am

“Mining for Emerging Technologies within Text Streams and Documents”

Dave Engel, Paul Whitney, Gus Calapristi, Fred Brockman

9:30-10:00am

“The Role of Semantic History on Online Generative Topic Modeling”

Loulwah AlSumait, Daniel Barbara, Carlotta Domeniconi

10:00-10:30am

Break

Session II:

Anomaly and Trend Detection

10:30-11:00am

“Threshold Setting and Performance Monitoring for Novel Text Mining”

Wenyin Tang, Flora S. Tsai

11:00-11:30am

“FutureLens: Software for Text Visualization and Tracking”

Gregory Shutt, *Andrey A. Puretskiy*, Michael W. Berry

11:30-12:00am

“Chatcoder: Toward the Tracking and Categorization of Internet Predators ”

April Kontostathis, Lynne Edwards, Amanda Leatherman

12:00-1:15pm

Lunch

Session III:

Text Classification and Clustering

1:30-2:00pm

“Content-based Spam Filtering by Machine Learning Algorithms”
Eric Jiang

2:00-2:30pm

“E-Mail Classification Based on NMF”
Andreas Janecek, Wilfried Gansterer

2:30-3:00pm

“Constrained clustering with k -means”
Ziqiu Su, Jacob Kogan, Charles Nicholas