Introduction to Web Mining Minitrack in Internet and Digital Economy Track

Dave King
JDA Software
Daveking63@gmail.com

Web Mining is the application of data mining and information extraction techniques aimed at discovering patterns and knowledge from the Web. Traditionally, Web Mining is divided into three classes:

- **Web Content Mining** – discovery of useful information from text, image, audio or video data in the Web.
- **Web Structure Mining** – analysis of the node and connection (graph) structure underlying single web sites, as well as larger collections of interrelated sites.
- **Web Usage Mining** – often called Web analytics involves extracting useful information from server logs and other sources detailing usage patterns.

Examples of more specific topics of interest include but are not limited to the following:

- Text mining of Web and Social Media content
- Opinion mining and sentiment analysis
- Web usage analysis
- Link analysis
- Analysis of search behavior
- Predictive analytics based on Web and social media content and search behavior
- Recommendation analysis
- Visual analysis of Web structure, usage, and content
- Semantic representations of Web content and linkages
- Analysis of Web-based collective intelligence

This year three papers have been accepted to the minitrack including:

- **Predicting the Viral Web** by Muhammad Gulzar et al. Provides a framework for detecting top trending topics and the length of those topics and demonstrates the utility of the framework by applying it to health related discussion forums.
- **Text-Generated Fashion Influence Model: An Empirical Study on Style.com** by Yusan Lin et al. Offers a novel approach for analyzing the influence of leading fashion designers by examining the designers influence network utilizing data from 6000+ runway reviews collected from Style.com from 2000-2014.
- **USpam -- A User Centric Ontology Driven Spam Detection System** by Muhammad Shoaib and Muddassar Farooq. Offers a new spam detection system based on the use of ontologies to model features that are extracted from a users profile. The features are then employed by machine learning classifiers - J48 and Naive Bayes - to make judgments about the relevance to the user's interests. The system called USpam is test on ENRON's Spam datasets, resulting in a major reduction in the number of "false alarms."