

## Data Mining and Information Retrieval

H. Michael Chung  
California State University, Long Beach  
[hmchung@csulb.edu](mailto:hmchung@csulb.edu)

Fred Gey  
University of California, Berkeley  
[gey@usdata.berkeley.edu](mailto:gey@usdata.berkeley.edu)

Selwyn Piramuthu  
University of Florida, Gainesville  
[selwyn@nersp.nerdc.ufl.edu](mailto:selwyn@nersp.nerdc.ufl.edu)

This minitrack covers the broad theory and application issues related to data mining, machine learning, knowledge acquisition, knowledge discovery, information retrieval, database, and inductive decision-making. Both structured and unstructured data repositories including human expert decisions, environmental/normative datasets, large document collections, and web databases are considered. Theoretical and methodological exploration in the previous years motivates us to further investigate the various and richer data and knowledge representation schemes such as Web, multimedia, and geographic data applied to science as well as management domains.

Information Retrieval algorithms support the computerized search of large document collections (millions of documents) to retrieval small subsets of documents relevant to a user's information need. Such algorithms are the basis for Internet search engines and digital library catalogues.

Based on this philosophical direction, we accepted four papers to be presented in this session. The first two papers deal with information/text retrieval methods. Both papers deal with Web-based retrieval methods and, interestingly enough, both draw their experimental basis from the Open Directory Project, the largest human edited index of the WWW, developed and maintained by volunteers (see <http://www.dmoz.org>). The third paper deals with load balancing in distributed file system

servers based on mining access patterns. In this paper, Glagoleva and Sathaye present a tool to perform load balancing across multiple servers. The purpose here is to manage "read-write" file sets using association rule discovery and graph analysis algorithms. The fourth presentation is on the characteristics of data used as input to decision tree generators. Piramuthu evaluates the effects of a few different characteristics and their modifications on the performance of decision trees.

The focus of the paper by Monroe, French and Powell is distributed collection searching by domain, and how sub-domain collections can be characterized by sampling of vocabulary present within html pages. Their experimental results seem to indicate that starting with a seed query word and iteratively sampling of resulting html pages for subsequent seed query words leads to 'good enough' vocabulary convergence after about 500 pages are sampled within a domain area.

The paper by Tanudjaja and Mui presents a graph-theoretic model using link-based retrieval in combination with personalized profile of the user's vocabulary to raise or 'lift' individual nodes from within a cluster of documents. The theory is folded into a 'personalization' module to wrap around a search engine. Different methods of lifting are tested by simulation experiments.

Glagoleva and Sathaye present a web based distributed file system server management tool to perform load

balancing across multiple servers. The proposed method generates association rules identifying inherent DFS file access patterns. Using real world data, they show improved performance using the proposed method.

The talk by Piramuthu is on input data characteristics for decision trees and their effects on learned concepts. Traditional statistical regression analysis assumes certain distribution (e.g., Gaussian) of input data, as well as other characteristics of data such as the data being independent and identically distributed. In most real world data, some of these assumptions are often violated. And, there are several means to at least partially rectify some of the consequences that arise from these violations.

This presentation considers a few of these situations: non-linearity in the data, the presence of outliers in the data, the presence of heteroschedasticity in the data, and the presence of multicollinearity in the data. See-5.0 is used as the decision tree generator to study the effects of these situations on input data for decision trees.

Preliminary results indicate that the performance of decision trees can be improved by considering the effects due to non-linearity, outliers, heteroschedasticity, and multicollinearity in input data. Both non-linearity and the presence of outliers did affect the classification performance of decision trees. The presence of heteroschedasticity did not affect the classification performance of decision trees significantly. And, the presence of multicollinearity is not of concern for decision trees. The attempt to remove multicollinearity resulted in poor classification performance.