

Data Mining and Information Retrieval

Selwyn Piramuthu
University of Florida, Gainesville
selwyn@ufl.edu

H. Michael Chung
California State University, Long Beach
hmchung@csulb.edu

The minitrack covers the broad theory and application issues related to data mining, machine learning, knowledge acquisition, knowledge discovery, information retrieval, data base, 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 computerized search of large document collections to retrieve 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 selected six papers to be presented in this session. The paper by Beierle and Kern-Isberner consider knowledge discovery as an inverse operation to inductive knowledge representation. This allows them to precisely define relevance with respect to a given inductive representation method. They use several examples to illustrate the method.

The paper by Lin and Lee deals with incremental sequential pattern mining. The proposed method (KISP) generates a knowledge base accumulating pattern information from individual mining, and speeds up the entire process of sequential pattern mining. They also generate the reduced candidate sets directly and count variable sized candidates concurrently. They compare the proposed method with GSP, a sequence mining algorithm.

The paper by Lin and Hsueh utilizes information retrieval and data mining techniques to develop approaches for knowledge map creation and maintenance in virtual communities of practice. They use both synthetic and real-world data evaluated by domain experts to illustrate the proposed method. They claim that this method enables dynamic knowledge management of professional communities on the Internet.

Natter and Mild present DELI, an interactive tool for new product development and targeting. They claim that their approach is able to cope with a large number of potential product features and levels unlike existing market research methods such as conjoint analysis. Its main advantages lie in the integration of segmentation, visualization of competitive structures, segment specific identification of new product functionality, and several interactive features that support the search for new products.

The paper by Damianos, Wohlever, Kozierek, and Ponte presents a case study of integrated knowledge discovery tools using MiTAP, a system that was developed as an experimental prototype using human language technologies for monitoring infectious disease outbreaks. They attribute the success of this system to its user-focused design that accommodates imperfect component technologies and its interactive features. They discuss this system in detail in this paper.

The paper by Kloptchenko, Back, Vanharanta, Toivonen, and Visa presents a method for allocating conference papers using prototype matching. They use data from past HICSS meetings to evaluate their method.