

Information Retrieval and Digital Library Applications Minitrack

Ray R. Larson, co-chair
*School of Information Management and
Systems*
University of California, Berkeley
ray@sims.berkeley.edu

Fredric C. Gey, co-chair
U.C. Data Archive & Technical Assistance
University of California, Berkeley
gey@ucdata.berkeley.edu

Over the past decade, since the rise of internet search engines and fact-based question-answering, the computer science research field of information retrieval has become more widely known and has attracted a number of researchers from other areas of computer science. Digital libraries are a somewhat newer specialty within computer and information science that focus on provision and access to large-scale organized collections of digital materials.

Information Retrieval algorithms support the computerized search of large document collections (millions of documents) to retrieve small subsets of documents estimated to be relevant to a user's stated information need. Such algorithms are the basis for internet search engines and digital library catalogues. The fundamental models for retrieval include Boolean logic (and its generalization through probabilistic "fuzzy logic"), geometric/vector space similarity, and probabilistic document retrieval.

Application areas within the broad fields of Information Retrieval (IR) and Digital Libraries (DL) include foreign language and cross-language retrieval, text categorization, text summarization, speech and broadcast, multimedia and image content retrieval. One characteristic of IR and DL research has been an emphasis on experimental validation and evaluation of the performance of different IR algorithms, typically such performance evaluation is performed against test collections of hundreds of queries matched to millions of documents.

This minitrack will examine some of the theoretical and application issues related to information retrieval, cross-language document search, link-based web search, text summarization, and fact-based question-answering as well as the applications of these technologies in Digital Libraries. The minitrack builds upon the section of Information Retrieval in Data Mining and Information Retrieval of HICSS-35 and the Digital Libraries and Information Retrieval minitracks within the Digital Documents and Media Track in 2003. The minitrack this year includes papers that examine a wide range of areas with the broad fields of information retrieval and digital libraries. This minitrack includes five accepted papers and a plenary lecture.

The topics examined in the accepted papers reflect the wide range of interests and research within this field. Uzuner, Katz and Davis examine methods of comparing the content, and expressions of that content, in documents. Methods and needs for preservation of digital objects over time as technology and standards change are discussed by Heminger and Kelley. Maybury and Griffith describe an Integrated environment for examining very large multilingual collections of data providing an environment for information analysts. Two of the papers examine the use of Genetic algorithms to different aspects of information retrieval and digital libraries. Wu and Agogino discuss how genetic algorithms have been used to extract descriptive terms from documents. Genetic algorithms are also used to discover optimal retrieval ranking functions for information retrieval on the WWW by Fan, Gordon, Pathak, Xi, and Fox.

In addition, this minitrack is sponsoring a plenary lecture on current research in multi-lingual text summarization, which will be presented by Dr. Judith Klavans of Columbia University. Dr. Klavans will discuss the **Columbia Newsblaster** system which summarizes news from over 40 www English language news sites and 25 foreign language sites in six languages including French, Spanish, Italian, German, Japanese, and Russian.