

Understanding and Communication

Michael Shepherd
Faculty of Computer Science
Dalhousie University
Halifax, Nova Scotia, Canada B3J 1W5
shepherd@cs.dal.ca

James W. Cooper
IBM T.J. Watson Research Center
P.O. Box 704
Yorktown Heights, NY 10598
jwcnmr@watson.ibm.com

The explosion of digital documents on the Internet and in the workplace has led to an increasing need for computer systems that help us not only manage the documents but also manage our understanding of these documents and their relationships. These digital documents include speech documents, and video and images as well as text documents in digital form.

This minitrack focuses on how one gains an understanding of a digital document and how that information is communicated. It encompasses retrieval and text analysis methods, including summarization, categorization, genre theory and detection, Web navigation and visualization methods that increase understanding of document content and genre.

This is the second year of this minitrack which is the result of merging two previous successful minitracks, Understanding and Visualization and Genre in Digital Documents. As such, there are continuing themes from both of the previous minitracks and from last year's successful minitrack.

This year there are two sessions. The first session has papers by Boongoen et al. and by Rehm. Both these papers attempt to "understand" documents but from very different approaches. Boongoen uses a network of agents and a natural language approach to extract knowledge from textual sources. The process is evolutionary in nature in that context gained from one document is used to help interpret the next document. Rehm's approach is to identify the Web genre of documents. His approach is to view a genre as consisting of a set of sub-genres which can then be automatically identified and extracted from the main genre.

Also in this session, Cooper et al. have developed a novel method for detecting similar documents through text mining techniques. These techniques can be used when several versions of the same document occur on various servers, documents are in different forms, e.g., HTML and PDF, and when one document is embedded in another. The results from this are very promising.

In the second session, Jones et al. continue the theme of document summarization using phrase extraction. The algorithm supports the dynamic summary resizing and refocusing facilities provided through the user interface and an evaluation of the system indicates better results than other baseline measures.

Shepherd et al. return to the problem of filtering of electronic news. In this machine learning approach, stereotypes are combined with neural nets to develop individual profiles. Two types of tasks were examined, the task where there is no explicit information need and the corporate profile where there is an explicit information need. Once again, it was found to be virtually impossible to filter news when the task is ludic in nature and almost impossible when the task is more focused.

The final paper in this session is quite different from previous themes. Spangler et al. apply multiple taxonomies and visualization in an attempt to understand a document collection. First they generate multiple taxonomies from the corpus then they provide a radial graph from which the user can select a particular class of documents to examine. Once a class is selected, the user can refine the query from tools presented and eventually the entire class is mapped and presented to the user visually.