DIGITAL DOCUMENT UNDERSTANDING AND VISUALIZATION

James W. Cooper
IBM Thomas J Watson Research Center

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.

When users are confronted with hard copy documents or unorganized digital documents, they look at main topic headers, skim the contents and so forth. The problem of how this can be done more effectively in an on-line environment is the main topic of this minitrack. Accordingly, this minitrack brings together the multiplicity of research in various fields on how organizations, groups and individual users seek to understand and navigate through document collections and individual documents.

Possible areas for papers included summarization, categorization, and key-phrase clustering, user interfaces for understanding documents, and studies of how users access documents and intuit their contents without actually reading much of them.

In this session, Kazi and Ravin tackle the problem of resolving ambiguity in recognizing who or what a name or term refers across a collection of documents. This is valuable in designing methods of summarizing the contents of a document collection as well as in recognizing related documents.

Boguraev and Neff describe approaches to the problem of recognizing segments in longer documents and determining topic shifts. Solving this problem can help produce more accurate summaries of multi-topic articles.

Colbath, et. al. describe an approach to searching archives of spoken documents. This is a particularly difficult problem because of the lack of formatting cues and the inherent inaccuracy of transcribed speech.

Roussinov and McQuaid present methods for clustering documents and investigate the efficacy of these techniques in actual user experiments.

This session also includes two proposals for work, and descriptions of the progress of these proposals. Futukata proposed how documents may be arranged based on observing the processes used to arrive at the generation of the document, and Wilde describes how documents can be organized into a link-based view of the collection.

Finally, Cooper and Prager describe a solution to the problem of eliminating documents of minimal content from search results and shows how document similarity can be used to validate these content-free documents.