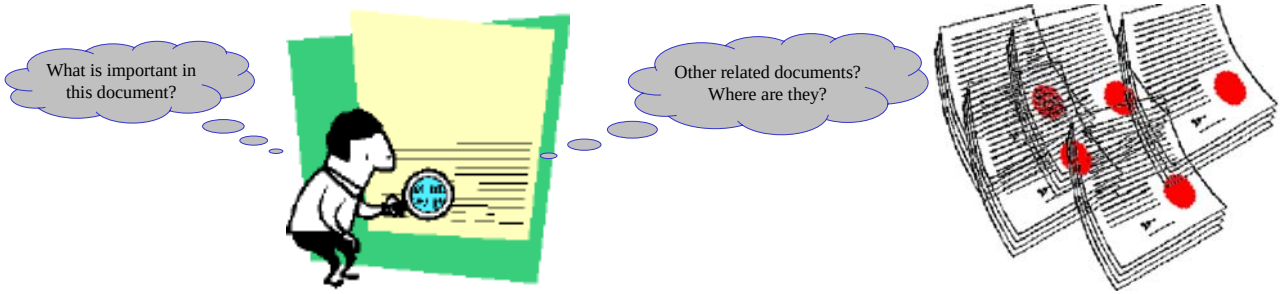
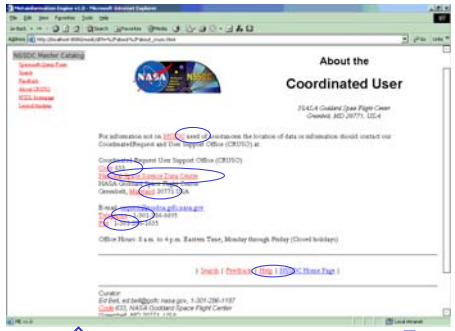
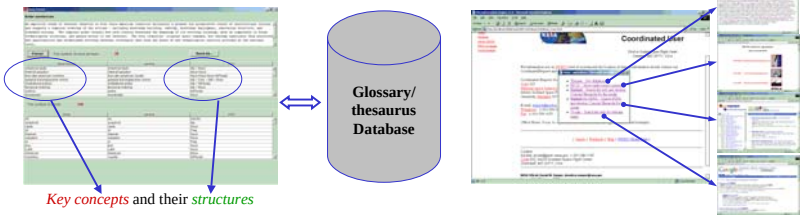
Lexical Analysis: Key-phrase Extraction and Concepts Organization

Brook Wu (wu@njit.edu) and Xin Chen (xc7@njit.edu), New Jersey Institute of Technology

This project is part of Digital Library Services Integration Project PI: Michael Bieberr, Co-PI: Brook Wu and Il Im







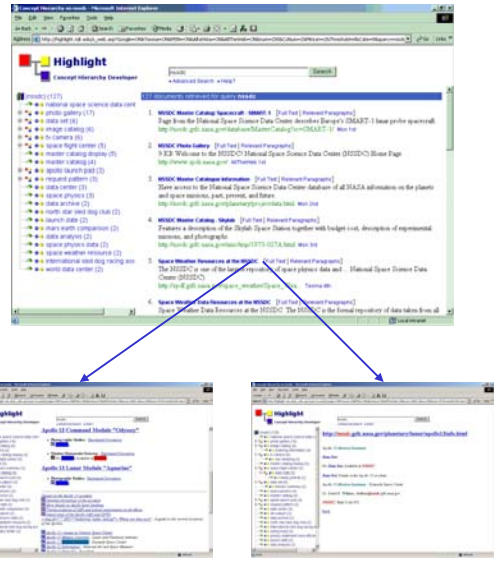
Key concepts and their *structures*

Noun phrases are extracted first. WordNet is used to assign the initial Part-Of-Speech for each word. Syntactic rules and heuristics are used to disambiguate words with multiple POS tags. Noun phrases are identified by applying POS pattern matching. Glossary/thesaurus database is used if only human identified concepts are of interests.

Automatic Key Phrase Extraction

Extracted concepts are associated with a list of links identified by relationship analysis. Each link leads to a particular service for the associated concept provided by other systems.

Resource Linking



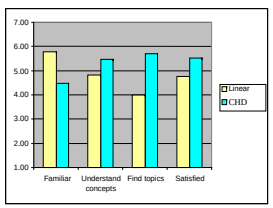
Dynamic Topical Highlight - Full Text vs. Relevant Paragraph Mode

Concept Hierarchy Developer uses indexed concept phrases and their co-occurrences in the text to develop the ***concept hierarchy*** on-the-fly. Upon selecting items of interest, users are provided with all relevant paragraphs in the returned document set. Users can also choose to read full text of a relevant document, if more information is desired.

Concepts Organization

User Study

- Though subjects are more familiar with the linear interface, they are more satisfied with the concept hierarchy interface.
- The concept hierarchy helps in finding more new topics related to the query.
- With the summarization hierarchy subjects understand the concepts in the document set better.
- Usefulness: 94% of the subjects found the topical highlighting useful; 74% of the subjects found the relevant paragraph highlighting useful.



Search Pattern

By analyzing the click sequence in the log files, we found that

- On the linear list interface, over 80% of the patterns were sequential i.e. documents were clicked sequentially as displayed.
- On the hierarchical interface, over 80% of the patterns were not sequential. It suggests that the subjects reach the documents based on their interests i.e. choosing different branches in the concept hierarchy.
- In over 85% of the cases, the lowest ranked document in the hierarchical interface was always significantly lower than that in the linear interface. Therefore, the hierarchical interface allows users to access documents of interest even if they are ranked lower.

Search Efficiency

Search efficiency is defined as the number of useful docs divided by search duration (minutes). With a p value < 0.05, the search efficiencies for the hierarchical interface were significantly better than that of its linear interface.

System Performance

Application environment: Intel Pentium 1700MHz, 512M RAM, MS Windows Advance Server 2000, MS IIS 5.0

Average searching time (200 documents in total from 5 web search engines): 2500ms

Average Concept Hierarchy development time: 80ms

Evaluation