Using Superimposed and Context Information to Find and Re-find Sub-documents

Sudarshan Murthy
Department of CS, Portland State University

Uma Murthy, Edward Fox
Department of CS, Virginia Tech
Sub-document Search

• Users frequently search for sub-documents, but search engines tend to return list of documents
  – Not much indication of why the returned documents are relevant, and which parts of a document are relevant
• Some search engines include snippets, but provide no assistance to identify sub-documents within a document
  – Users routinely open documents and search intra-document for relevant sub-documents
• Finding effort is often repeated when re-finding
Windows Desktop Search

Search term: environmental assessment

Documents returned:
Google Desktop

Search term

Documents returned
An Alternative: *Super Search*

• See list of documents and sub-documents along with detailed context information for sub-documents
  – Content: Sentence, paragraph, row, column, …
  – Location: Page number, line number, sheet name, …

• See sub-documents in context: Takes users *directly* to sub-documents; does not require authors to identify sub-document units

• Alternatively, browse context information without clicking through to sub-documents
Document Results

Search term

Documents returned
Sub-document Results*

* Sub-document results obtained using mark harvesters; context information obtained using the SPARCE context mechanism
See Detailed Context Information
See a Sub-document in Context
Re-finding Sub-documents

Information “found” earlier is ranked higher

Context information can help re-find information
Super Search Benefits

• Judge relevance more easily
  – See sub-documents, not just documents

• Reduce user click-through operations
  – Context information provides insight into sub-documents

• Reduce effort to find sub-documents in context
  – No need for user to perform *intra-document* search

• Leverage finding effort to re-find
  – See found results first; easy to remember found info.
Discussion Points

• Current search engine techniques can be inefficient when supporting finding and re-finding of sub-documents in this manner
  – Heterogeneous information
  – Showing information in context
  – How much of context information can a traditional inverted index provide?
    • Need to distinguish paragraphs, lines, rows, columns, page, …