Text and Image Retrieval in Cheshire II

Ray R. Larson
School of Information Management and Systems
University of California, Berkeley
510-642-6046
ray@sherlock.berkeley.edu

1. DEMONSTRATION DESCRIPTION

The Cheshire II system was originally designed to apply probabilistic retrieval methods to online library catalog searches in order to help overcome the pervasive twin problems of topical searching of Boolean online catalogs: search failure and information overload[1]. It was intended to be a next-generation online catalog and full-text information retrieval system that would apply probabilistic retrieval methods to simple MARC records and clustered record surrogates. Over time the system has been explanded to include support for full-text SGML documents (ranging from simple document types as used in the TREC database[2] to complex full-text document encoded using the TEI and EAD DTDs) and support for full-text OCR from scanned page image files linked to SGML bibliographic records. Recently the system has been used to provide fast access to natural images using the BlobWorld image segmenting system. The system is the primary text search engine for the UC Berkeley Environmental Digital Library project sponsored by NSF, NASA, and ARPA. It also provides access to a number of diverse databases databases via the WWW using an HTTP to Z39.50 gateway. It has also been adopted for use as a search engine in working library environments including the Physical Sciences Libraries at UC Berkeley, The Data Archives at the University of Essex, and the special collection department of the University of Liverpool Library. The Cheshire II system includes the following features:

1. It supports SGML/XML as the primary data base format of the underlying search engine, and provides support for full-text data linked to SGML metadata records. Traditional online catalog databases are supported using MARC to SGML conversion.
2. It is a client/server application where the interfaces (clients) communicate with the search engine (server) using the Z39.50 v.3 Information Retrieval Protocol. The system also provides a general Z39.50 Gateway with support for mapping Z39.50 queries to local Cheshire databases and to relational databases.
3. It includes a graphical direct manipulation interface using Tcl/Tk on unix and NT platforms, as well as a CGI interpreter version that combines client and server capabilities.
4. It allows users to enter queries as "free-text" (that is, normal English prose) statements of their interest or need. Full Boolean is also available.
5. It uses probabilistic ranking techniques to match the user's initial query with documents in the database. In some databases it can provide two-stage searching where a set of ``classification clusters'' is retrieved ranked by probable relevance to the user's search statement. These are then used for feedback about the primary topical areas of the query, and documents are retrieved based on the selected clusters. This aids the user in subject focusing and topic/treatment discrimination.
6. It supports open-ended, exploratory browsing through following dynamically established linkages between records in the database. These can be dynamically generated ``hypersearches'' that let users issue a Boolean query with a mouse click to find all items that share some field with a displayed record.
7. It uses the user's selection of relevant citations to refine the initial search statement and automatically construct new search statements for relevance feedback searching.

A primary goal of the Cheshire II system design was to provide an extensible system that can easily adapt to new types of data, and provide a flexible and programmable user interface to display that data. In order to achieve this goal, we have incorporated appropriate national and international standards into the system wherever possible.

Acknowledgements: Development of the Cheshire II system has been sponsored as part of Berkeley's NSF/NASA/ARPA Digital Library Initiative Grant#IRI-9411334. Current work is being sponsored by DARPA Contract N66001-97-C-8541; AO# F477.

2. REFERENCES

[1] Larson, R.R., McDonough, J., O'Leary, P., Kuntz, L. & Moon, R. Cheshire II: Designing a next-generation online catalog. *Journal of the American Society for Information Science* 47(7) (July 1996) 555-567.

[2] Larson, R. R & McDonough, J. Cheshire II at TREC 6: Interactive Probabilistic Retrieval. In E. Voorhees and D. Harman (Eds.) *Information Technology: The Sixth Text Retrieval Conference (TREC-6).* (pp. 649-649) Gaithersburg, MD : NIST, August 1998.