Berkeley's TREC 8 Interactive Track Entry: Cheshire II and Zprise

Ray R. Larson
School of Information Management and Systems
University of California, Berkeley
Berkeley, CA 94720-4600
ray@sherlock.berkeley.edu

Abstract

This paper briefly discusses the UC Berkeley entry in the TREC8 Interactive Track. In this year's study twelve searchers conducted six searches each, half on the Cheshire II system and the other half on the Zprise system, for a total of 72 searches. Questionnaires were administered to each participant to gather information about basic demographic and searching experience, about each search, about each of the systems, and finally, about the user's perceptions of the systems. In this paper I will briefly describe the systems used in the study and how they differ in design goals and implementation. The results of the interactive track evaluations and the information derived from the questionnaires are then discussed and future improvements to the Cheshire II system are considered.

Introduction

The primary goals of UC Berkeley entry in the TREC-8 Interactive track were to 1) attempt to replicate our entry in the TREC-6 and TREC-7 Interactive track with a larger number of participants (searchers), and 2) to evaluate changes to the experiment system (Cheshire II) to see if there were substantial differences in the ranking of the systems between previous year's entries and this year. In addition we are continuing to use the same systems, questionnaires, and complete TREC-7 Interactive track protocol to obtain further information that we hope to combine with the data obtained in previous TREC interactive track experiments for further analysis.

In TREC-8 we used virtually identical implementations of the Cheshire II system and the ZPRISE system as those used in previous TRECs. The database and indexing for each system were also the same as for TREC-6 and TREC-7 (Larson & McDonough, . The changes made to the Cheshire II system for this year's experiment are discussed below.

The Cheshire II System

The design and retrieval algorithm of the Cheshire II system have been discussed in both the TREC-6 and TREC-7 papers, and only the highlights of that description will be repeated here. The Cheshire II system finds its primary usage in full text or structured metadata collections based on SGML and XML, often as the search engine behind a variety of WWW-based "search pages" or as a Z39.50 server for particular applications. The Cheshire II system includes the following features:

1. It supports SGML and XML as the primary database format of the underlying search engine
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.
3. It includes a programmable graphical direct manipulation interface under X on Unix and NT. There is also CGI interpreter version that combines client and server capabilities.
4. It permits users to enter natural language queries and these may be combined with Boolean logic for users who wish to use it.

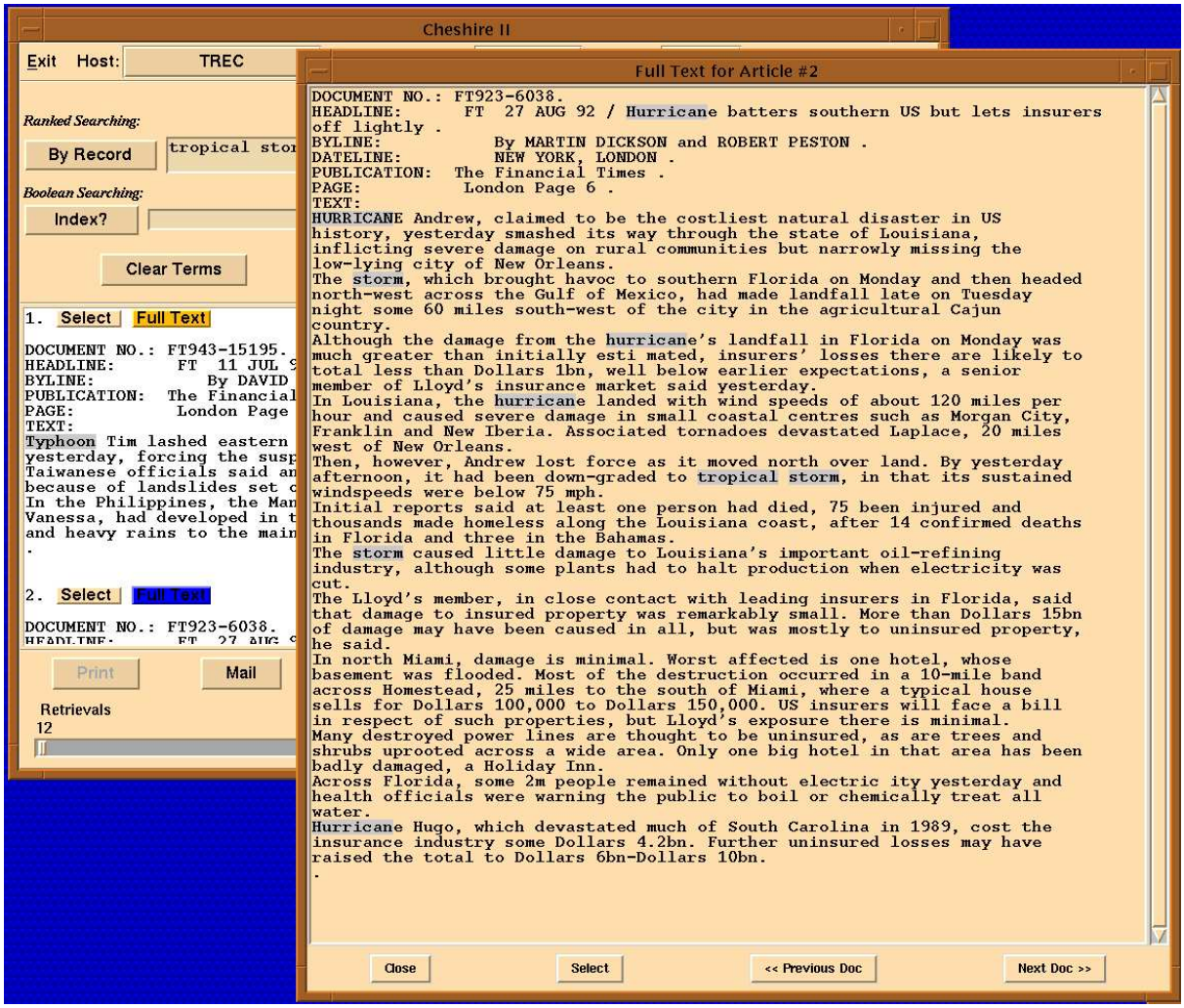

Figure 1: New Cheshire II Interface with Full-Text Window

5. It uses probabilistic ranking methods based on the Logistic Regression research carried out at Berkeley to match the user's initial query with documents in the database.
6. It supports open-ended, exploratory browsing through following dynamically established linkages between records in the database, in order to retrieve materials related to those already found. 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.

The Cheshire II search engine supports both probabilistic and Boolean searching. The design rationale and features of the Cheshire II search engine have been discussed in the TREC-6 and TREC-7 papers (Larson & McDonough, 1998; Gey, Jiang, Chen & Larson, 1999).

The Cheshire search engine functions as a Z39.50 information retrieval protocol server providing access to a set of databases. In the TREC-8 experiments the TREC Financial Times (FT) database was the only database used by participants. The system supports various methods for translating a searcher's query into the terms used in indexing the database. These methods include elimination of unused words using field-specific stopword lists, particular field-specific query-to-key conversion or "normalization" functions, standard stemming algorithms (Porter stemmer).

The Cheshire II search engine supports both Boolean and probabilistic searching on any indexed element of the database. In probabilistic searching, a natural language query can be used to retrieve the records that are estimated to have the highest probability of being relevant given the user's query. The search engine supports a simple form of relevance feedback, where any items found in an initial search (Boolean or probabilistic) can be selected and used as queries in a relevance feedback search.

The probabilistic retrieval algorithm used in the Cheshire II search engine is based on the *logistic regression* algorithms developed by Berkeley researchers (Cooper, et al. 1992, 1994a, 1994b). The Cheshire II search engine also supports complete Boolean operations on indexed elements in the database, and supports searches that combine probabilistic and Boolean elements.
Relevance feedback is supported and implemented quite simply, as probabilistic retrieval based on extraction of content-bearing elements (such as titles, subject headings, etc.) from any items that have already been seen and selected by a user. At the present time we do not use any methods for eliminating poor search terms from the selected records, nor special enhancements for terms common between multiple selected records (Salton & Buckley 1990).

The Cheshire II Client Interface

The design of the Cheshire II client interface (shown with the TREC FT database in Figure 1), has also been discussed in previous TREC papers. This discussion will concentrate on changes made to the interface for the purposes of our TREC-8 experiment. The Cheshire II interface was intended to provide a generic interface to Z39.50 servers, primarily for search and display of library catalog information and other bibliographic databases. The principle design goals in the interface design were:

1. to support a consistent interface to a wide variety of Z39.50 servers, and to dynamically adapt to the particular server.
2. to reduce the cognitive load on the users wishing to interact with multiple distributed information retrieval systems by providing a single interface for them all.
3. to minimize use of additional windows during users' interactions with the client in order to allow them to concentrate on formulating queries and evaluating the results, and not expend additional mental effort and time switching their focus of attention from the search interface to display clients;

As pointed out in the TREC-7 paper (Gey, Jiang, Chen & Larson, 1999), the interface design assumed that most of the information retrieved and viewed in the search interface would be brief metadata records for documents, and not full text documents themselves. The ability to view full-text documents such as the FT articles used in the interactive track experiments was initially added to the existing interface as longer records that could be scrolled in the main display window. However, comments and questionnaire responses from TREC-7 participants indicated that the separate document viewing window associated with the ZPRISE system was preferable to having to do so much scrolling to accomplish the Interactive Track tasks..
The primary addition to the Cheshire II client interface was the addition of a full-text display window that included controls for selecting/saving the displayed document. This window is shown in Figure 1. The full-text window is invoked by the "Full Text" button next to the "Select" button for each record. The "Full

Text" button changes color to indicate the currently displayed full-text document (blue) or previously seen documents (orange/gold). The full-text window also included controls for stepping directly to the next or previous full-text document in the retrieval list.

In addition, the Boolean NOT, requested by several searchers in TREC-7 was brought out to the interface and integrated with the Boolean search capability.

The Zprise System

The second (control) system used in the TREC-7 Interactive track at Berkeley was the Zprise system from NIST. This system was used in the same configuration and with the same database indexing setup as used for the global control system in our TREC-6 and TREC-7 Interactive Track entries. Zprise, as configured for this test was limited to a total of 24 retrieved items and relevance feedback was disabled. However, the interface was set up so that it provided a very good fit for the tasks involved in the interactive track. For example, documents were viewed in full text form in a separate window from the short display (consisting primarily of title and date as well as control elements for indicating relevant documents and for moving around in the brief display. Most of our users found the ZPRISE displays simple to learn and to operate, in fact most found that the operations required to carry out the Interactive Track tasks were easier to do on the ZPRISE interface than they were on the Cheshire II interface. This was not entirely surprising, since the ZPRISE interface is designed to support TREC-like databases containing full text. We had hoped that the addition of the full-text display to the Cheshire II system would show less difference in preference (and hopefully, less differences in the aspectual recall and precision figures) when compared to TREC-7. But, as discussed below, this hope was not fulfilled.

TREC Interactive Track

The administration of the Interactive Track followed the protocols set down in the track guidelines. This mandated a minimum group of 12 participant searchers, each of whom conduct 6 searches, half on the control system (ZPRISE, identified as "Z") and half on the experimental system (Cheshire II, identified as "C"). Each searcher was asked to use the features of the respective interfaces to select as relevant those documents that they considered to relevant to one or more aspects of the specific topic.

		Topic						
System	Data	408i	414i	428i	431i	438i	446i	Overall Average
C	Average of Recall	0.39	0.72	0.37	0.37	0.21	0.24	**0.38**
	Average of Precision	0.80	0.61	0.77	0.84	0.67	0.44	**0.69**
Z	Average of Recall	0.42	0.72	0.42	0.40	0.21	0.29	**0.41**
	Average of Precision	0.93	0.59	0.68	0.86	0.86	0.75	**0.78**

Table 1. Average Precision and Recall by Topic for Cheshire and Zprise

The pooled results for all systems were evaluated at NIST by the TREC evaluators and "Aspectual Precision" and "Aspectual Recall" for each searcher was calculated. Table 1 shows the values for Aspectual Precision and Recall by TREC topic for the two Berkeley systems ("C" and "Z", the Cheshire II system and ZPRISE systems respectively) are shown in boldface in Tables 1 and 2. The control system "Z" performed

considerably better than the experimental system in terms of the Aspectual Precision and noticeably better in terms of Aspectual recall. Needless to say, this is a disappointing result, and our analysis has yet to reveal any obvious reason for the discrepancy. We believe that the difference may be due to the more complex interactions required to perform the search tasks on the generic Cheshire II interface than on the ZPRISE system, certainly the comments of participants on the questionnaires indicated that most of them preferred the ZPRISE system.

		System		
Searcher	Data	C	Z	Mean
P1	Average of easy_start	3.6667	4.6667	4.1667
	Average of easy_search	4.3333	4.3333	4.3333
P10	Average of easy_start	5.0000	4.6667	4.8333
	Average of easy_search	4.6667	4.6667	4.6667
P11	Average of easy_start	3.6667	3.3333	3.5000
	Average of easy_search	3.6667	3.6667	3.6667
P12	Average of easy_start	3.0000	3.6667	3.3333
	Average of easy_search	3.6667	3.6667	3.6667
P2	Average of easy_start	3.0000	3.6667	3.3333
	Average of easy_search	2.3333	3.0000	2.6667
P3	Average of easy_start	3.3333	3.0000	3.1667
	Average of easy_search	3.6667	3.3333	3.5000
P4	Average of easy_start	3.0000	3.6667	3.3333
	Average of easy_search	3.0000	3.3333	3.1667
P5	Average of easy_start	2.3333	3.6667	3.0000
	Average of easy_search	2.0000	3.0000	2.5000
P6	Average of easy_start	3.0000	3.3333	3.1667
	Average of easy_search	3.0000	3.3333	3.1667
P7	Average of easy_start	4.0000	4.6667	4.3333
	Average of easy_search	4.0000	4.3333	4.1667
P8	Average of easy_start	3.3333	4.0000	3.6667
	Average of easy_search	3.3333	4.3333	3.8333
P9	Average of easy_start	4.0000	4.0000	4.0000
	Average of easy_search	3.6667	4.0000	3.8333
Mean of “easy to start searching”		3.4444	3.8611	3.6528
Mean of “easy to search”		3.4444	3.7500	3.5972

Table 3: Average Ease of Starting Search and Ease of Doing Search for each Participant by system

In the following section we will examine the characteristics of the searchers as reported in the questionnaires administered during the experiments. Figure 3 summarizes the average aspectual precision and recall for each of the systems participating in the TREC-7 Interactive Track.

User Characteristics

The administration of the interactive track followed the track guidelines with a single group of 12 participants. While only one of the participants had used either the experimental (Cheshire II) or control

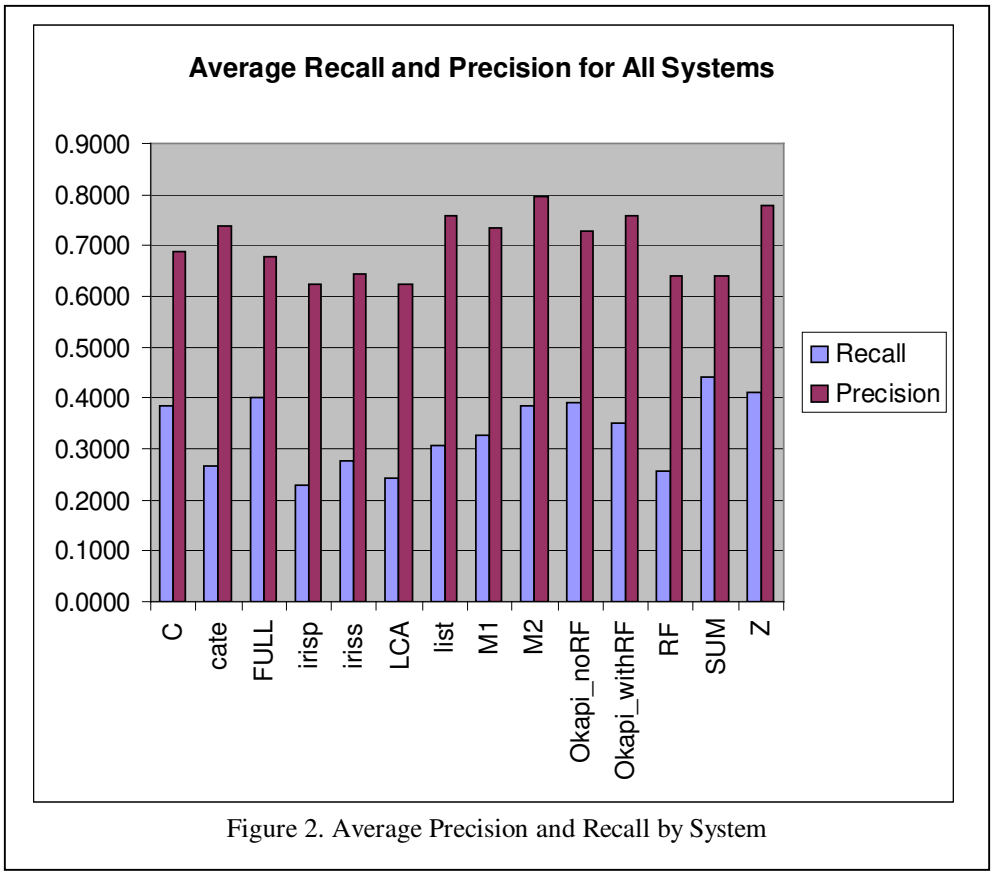

Figure 2. Average Precision and Recall by System

(ZPRISE) systems in searching tasks, some had seen demonstrations of the experimental system. The searchers who participated in the study were volunteers drawn from the School of Information Management and Systems at UC Berkeley (a call for participation was sent to all students and faculty at SIMS and the first 12 volunteers were scheduled for search sessions. A pre-search questionnaire asked each participant about:

1. What high school/college/univerity degrees/diplomas do have (or expect to have)?
2. What is your occupation?
3. What is your gender?
4. What is your age?
5. Have you participated in previous TREC searching studies?
6. Overall how long have you been doing online searching?
7. Experience with using a point-and-click interface (e.g. Windows, Macintosh)
8. Experience searching on computerized library catalogs either locally or remotely
9. Experience searching on CD-ROM systems
10. Experience searching on commercial online systems (BRS afterdark, Dialog, Lexis-Nexis, etc.)
11. Experience searching on the World Wide Web search services (Alta Vista, Excite, Yahoo, Hotbot, etc.)
12. Experience searching on other systems
13. How often do you conduct a search on any kind of system
14. "I enjoy carrying out information searches"

All of the participants, except one undergraduate, held college degrees (One held a PhD, Three others were PhD students with previous undergraduate and graduate degrees, and the remaining participants were Masters students in the SIMS program). Three of the participants (P1, P2, and P3) had over 8 years of experience in online searching on other systems. As observed last year, once again the most frequently used search systems were the Web search services and the next most frequent were online catalogs. It appears that most recent searchers will be gaining their experience from the WWW and possibly from online library catalogs, and will probably not have experience (or as much experience) with traditional Boolean systems such as Dialog.

Per Search Results

Following each search the participants were given a questionnaire asking:

1. Are you familiar with this topic
2. Was it easy to get started on this search
3. Was it easy to do the search on this topic
4. Are you satisfied with your search results
5. Are you confident that you identified all of the different instances for this topic
6. Did you have enough time to do an effective search.

Table 4 shows the average responses for the "easy to do the search" and "easy to get started on the search" questions by searcher and system. As may be seen from the table, many searchers found the search easier to do with the ZPRISE system than with the Cheshire II system. Similarly, Table 5 shows the average responses to the "Are you satisfied with the results" question. Here, the overall scores rate the searches done with the Cheshire II system slightly higher than for ZPRISE. Table 6 shows the average responses to the question "Are you familiar with this topic?" Here the responses show that the searchers where generally less familiar with the topics searched on the Cheshire system versus those on the ZPRISE system. Correlation analysis showed, however, no significant correlation between familiarity with a topic and either the ease of searching or the satifaction with search results.

Post-System Questions

The searches were conducted in blocks of 4 questions on each system. Following the searcher's interaction with a system, a post-system questionnaire was administered. This post-system questionnaire asked each

Average of satisfied	System		
Searcher	C	Z	Meanl
P1	4.6667	4.3333	4.5000
P10	4.3333	3.6667	4.0000
P11	3.3333	3.0000	3.1667
P12	3.0000	3.0000	3.0000
P2	3.3333	2.3333	2.8333
P3	2.6667	2.3333	2.5000
P4	2.6667	2.6667	2.6667
P5	2.6667	3.0000	2.8333
P6	3.0000	3.3333	3.1667
P7	3.0000	3.3333	3.1667
P8	3.6667	4.6667	4.1667
P9	3.6667	4.0000	3.8333
Overall means	3.3333	3.3056	3.3194

Table 5: Average User Satisfaction with Search

Average of familiar	System		
Searcher	C	Z	Mean
P1	1.6667	2.0000	1.8333
P10	1.6667	3.6667	2.6667
P11	1.0000	1.3333	1.1667
P12	2.0000	1.6667	1.8333
P2	1.0000	1.6667	1.3333
P3	2.6667	2.0000	2.3333
P4	2.3333	1.6667	2.0000
P5	2.0000	2.3333	2.1667
P6	2.3333	2.3333	2.3333
P7	2.0000	3.0000	2.5000
P8	1.3333	2.6667	2.0000
P9	4.0000	4.0000	4.0000
Overall means	2.0000	2.3611	2.1806

Table 6: Average User Familiarity with Topics

searcher the following questions:

1. How easy was it to *learn to use* this information system?
2. How easy was it to *use* this information system?
3. How well did you *understand how to use* the information system?
4. Write down any comments that you have about your searching experience with this information retrieval system.

Overall, the searchers found both systems very easy to learn. The Cheshire system was marked down again on the "easy to use" question. From the comments, this appeared to be related to some features being hard to understand and use. Some searchers mentioned that it was hard to figure out when and if the items they selected as relevant had been seen before, and as previously observed, the need to scroll back to the beginning of a record to select it as relevant (for those NOT using the full-text window) was a problem when the full text is displayed in the main window.

Exit Questionnaire

After the completion of all searches an exit questionnaire was administered to the searchers. This questionnaire asked:

1. To what extent did you understand the nature of the searching task?
2. To what extent did you find this task similar to other searching tasks that you typically perform?
3. How different did you find the systems from one another?
4. Please rank the two systems in order of how *easy they were to learn to use.*
5. Please rank the two systems in order of how *easy they were to use.*
6. Please rank the two systems in the order of *which system you liked best.*
7. What did you like about each of the systems.
8. What did you dislike about each of the systems.
9. Please list any other comments that you have about your overall search experience.

The searchers claimed to have a very good understanding of the search task (mean was 4.16), and they found the task similar to other searching tasks (mean of 3.50). They also found the systems somewhat different (mean of 3.41). In ranking the systems, 7 out of 12 ranked Cheshire II as easier to learn to use, but only 5 out of 12 ranked it as easier to use. 7 out of the 12 searchers "liked" Cheshire the best of the two systems. However, as the Precision and Recall results show, they did not perform as well using the Cheshire system as they did using ZPRISE. One had a strong preference for the ZPRISE system, but commented that he might have preferred Cheshire if it had been introduced first.

Conclusions

It is very difficult to draw any firm conclusions from the analysis that we have conducted. There is no clear evidence why the Cheshire II system has shown poorer Precision and Recall performance that the control system. One tentative thought is that Cheshire II is providing *too much* functionality and may be confusing the users with too many options. Many of the users did use the Boolean features of the system, and this might have caused a significant reduction in Recall compared to the ranked retrieval offered by the ZPRISE system. These tentative hypotheses will need further analysis to discover if they are supported by the data collected.

Acknowledgements

I would like to thank SIMS PhD students Youngin Kim and Jacek Purat for their much needed help in conducting the user evaluation sessions for this research.

The original development of the Cheshire II system was sponsored by a College Library Technology and Cooperation Grants Program, HEA-IIA, Research and Demonstration Grant #R197D30040 from the U.S. Department of Education. Further development work on the Cheshire II project and system was supported as part of Berkeley's NSF/NASA/ARPA Digital Library Initiative Grant #IRI-9411334. Current work is being supported as part of the "Search Support for Unfamiliar Metadata Vocabularies" research project at UC. Berkeley, sponsored by DARPA contract N66001-97-C-8541; AO# F477. Future development of the Cheshire system is being sponsored by the NSF/JISC International Digital Libraries program.

Bibliography

Cooper, W. S., Gey, F. C., & Dabney, D. P. (1992). Probabilistic Retrieval Based on Staged Logistic Regression. In: *SIGIR '92 (Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Copenhagen, Denmark, June 21-24, 1992)* (pp. 198-210). New York: ACM.

Cooper, W. S., Gey, F. C. & Chen, A. (1994a). Full Text Retrieval based on a Probabilistic Equation with Coefficients fitted by Logistic Regression. In: D. K. Harman (Ed.) *Second Text Retrieval Conference (TREC-2), Gaithersburg, MD, USA, 31 Aug.-2 Sept. 1993, NIST-SP 500-215*, (pp. 57-66). Washington : NIST.

Cooper, W. S., Chen, A. & Gey, F. C. (1994b). Experiments in the Probabilistic Retrieval of Full Text Documents In: *Text Retrieval Conference (TREC-3) Draft Conference Papers, Gaithersburg, MD : National Institute of Standards and Technology.*

Gey, F. C., Jiang, H., Chen, A. & Larson, R. R. (1999) Manual Queries and Machine Translation in Cross-Language Retrieval and Interactive Retrieval with Cheshire II at TREC-7. In E. Voorhees and D. Harman (Eds.) *Information Technology: The Seventh Text Retrieval Conference (TREC-7).* NIST special publication 500-242. (pp. 527-540) Gaithersburg, MD : NIST, July 1999.

Larson, R. R. (1991). Classification Clustering, Probabilistic Information Retrieval, and the Online Catalog. *Library Quarterly*, 61, 133-173.

Larson, R. R. (1992). Evaluation of Advanced Retrieval Techniques in an Experimental Online Catalog. *Journal of the American Society for Information Science*, 43, 34-53.

Larson, R. R & McDonough, J (1998) Cheshire II at TREC 6: Interactive Probabilistic Retrieval. In E. Voorhees and D. Harman (Eds.) *Information Technology: The Sixth Text Retrieval Conference (TREC-6).* NIST special publication 500-240. (pp. 649-649) Gaithersburg, MD : NIST, August 1998.

Ousterhout, J. K. (1994). *Tcl and the Tk Toolkit* Reading, Mass. : Addison-Wesley.

Salton, G. & Buckley, C. (1990). Improving Retrieval Performance by Relevance Feedback. *Journal of the American Society for Information Science*, 41, 288-297.