Interactive Probabilistic Retrieval: Cheshire II at TREC 7

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 TREC7 Interactive Track. In this year's study eight searchers conducted eight searches each, half on the Cheshire II system and the other half on the Zprise system, for a total of 64 searches. Questionnaires were administered 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-7 Interactive track were to 1) attempt to replicate our entry in the TREC-6 Interactive track with a larger number of participants (searchers), and to see if there were substantial differences in the ranking of the systems between last year and this year, and 2) to follow the complete TREC-7 Interactive track protocol to obtain further information than obtained in TREC-7 via the standard questionnaires filled in by all searchers on all systems. We are hoping to develop a baseline that can be used to evaluate changes and additions to the systems (primarily to the Cheshire II system) in the future.

In TREC-7 we used virtually identical implementations of the Cheshire II system and the ZPRISE system as those used in TREC-6. The database and indexing for each system were also the same as for TREC-6. The characteristics of the Cheshire II system and ZPRISE systems are discussed below.

The Cheshire II System

The original design goals of the Cheshire II system were to develop a "next-generation" online library catalog system that would incorporate ranked retrieval based on probabilistic retrieval methods along with the Boolean retrieval expected in "second generation" online catalog systems. Much has changed since these initial goals were formulated. The Cheshire II system now 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. Current applications of the Cheshire II system include:

1. Use as a web-based library catalog for the UC Berkeley Physical Sciences libraries.
2. Use as the primary search engine for UC Berkeley NSF/NASA/ARPA Digital Library Project (where the system is used for searching full-text OCR from scanned page image files linked to SGML bibliographic records, as well as supporting image content retrieval as part of the "Blobworld" research effort.

3. Use as the search engine for the History Data Service at the University of Essex in the UK, as part of the Z3.50-based Arts and Humanities Data Service (AHDS).
4. Use as the search engine for collections of archival documents (in the Encoded Archival Description DTD) at the University of Liverpool in the UK.
5. It is being adopted as the metadata search engine for the Networked Social Science Tools and Resources (NESSTAR) project based at The Data Archive (Essex University) in the UK, the Danish Data Archives and Norwegian Social Science Data Services.
6. Use for California Sheet Music Project at UC Berkeley, linking MARC metadata with page images of early California music publications.
7. Use in the DARPA-sponsored Unfamiliar Metadata Vocabularies project at UC Berkeley where it is being used for probabilistic cross-language entry vocabulary mapping.

The Cheshire II system includes the following features:

1. It supports SGML and XML as the primary database format of the underlying search engine, and provides support for full-text data linked to SGML or XML metadata records. MARC format records for traditional online catalog databases are supported using MARC to SGML conversion software developed for the project.
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 programmable graphical direct manipulation interface under X on Unix with support for NT available in a pre-release form. There is also CGI interpreter version that combines client and server capabilities. These interfaces permit searches of the Cheshire II search engine as well as any other z39.50 compatible search engine on the network.
4. It permits users to enter natural language queries and these may be combined with Boolean logic for users who wish to use it.
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. In some databases it can provide two-stage searching where a set of "classification clusters"(Larson 1991) for the database is first retrieved in decreasing order of probable relevance to the user's search statement. These can then be used to provide feedback about the primary topical areas of the query, and retrieve documents within the topical area of 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, 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

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 paper (Larson & McDonough, 1997) and will only be briefly repeated here.

The search engine functions as a Z39.50 information retrieval protocol server providing access to a set of databases. In the TREC-7 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) and support for mapping database and query text words to single forms based on the WordNet dictionary and thesaurus using a adaption of the WordNet "Morphing" algorithm and

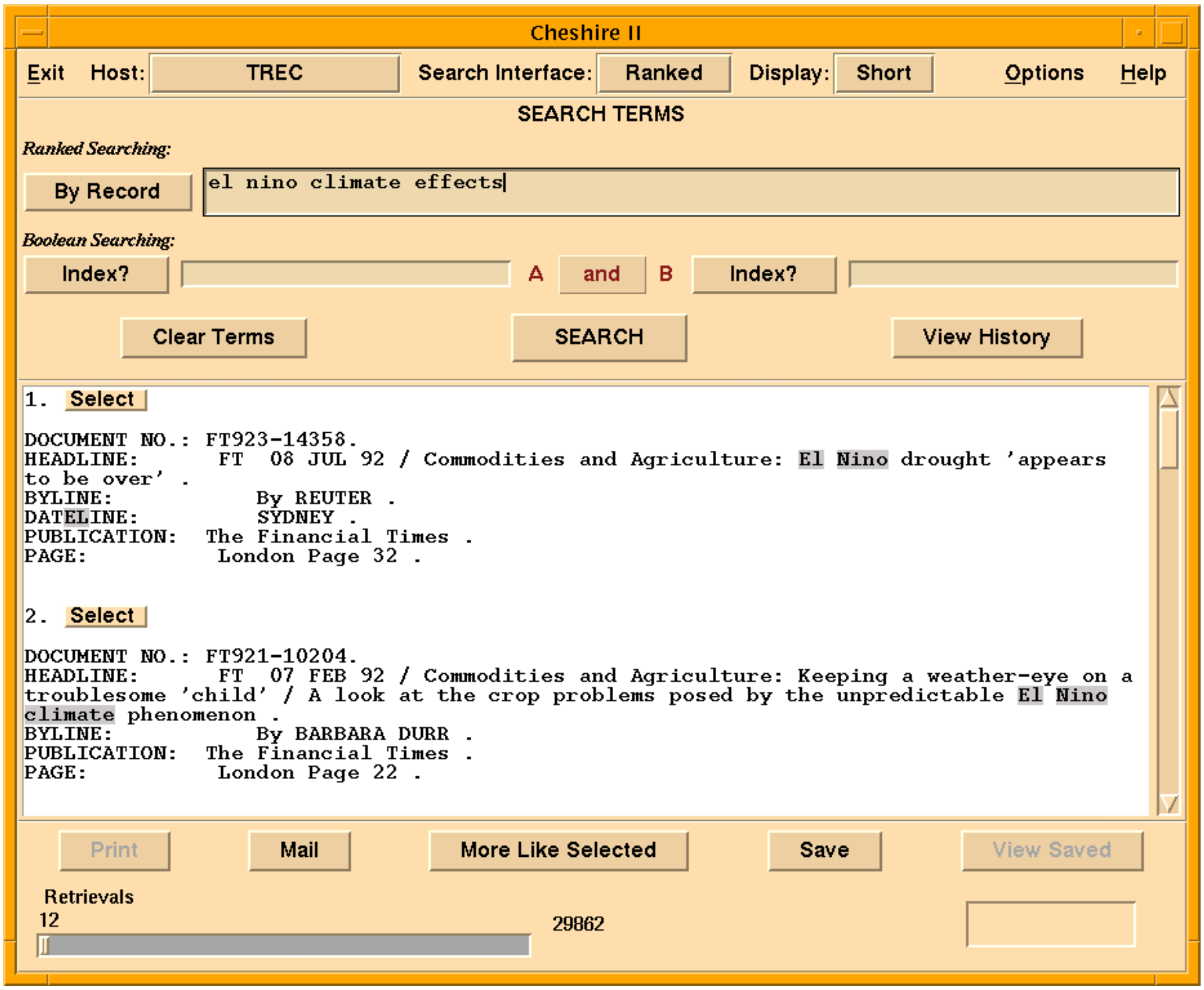

Figure 1. Cheshire II Short Display

exception dictionary..

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). Formally, the probability of relevance given a particular query and a particular record in the database $P(R \mid Q, D)$ is calculated and the records are presented to the user ranked in order of decreasing values of that probability. In the Cheshire II system P(R | Q, D) is calculated as the "log odds" of relevance log O(R | Q, D), where for any events A and B the odds O(A |B) is a simple transformation of the probabilities P(A | B) / P(~A | B). The Logistic Regression method provides estimates for a set of coefficients, c_i, associated with a set of S statistics, X_i, derived from the query and database, such that

$$\log O(R \mid Q, D) \cong c_0 \sum_{i=1}^{S} c_i X_i$$

where c_0 is the intercept term of the regression, for the set of M *terms* (i.e., words, stems or phrases) that occur in both a particular query and a given document. The regression equation and coefficients used in the TREC-7 Interactive Track are the same as used in our TREC-6 entry. These are based on the TREC-3 Adhoc entry from Berkeley (Cooper, et al. 1994b) where the regression coefficients were estimated using relevance judgements from the TIPSTER test collection. The basic elements are

This is the log of the absolute frequency of occurrence for term t_j in the query averaged over the M

$$X_1 = \frac{1}{M}\sum_{j=1}^{M} \log QAF_{t_j}$$

terms in common between the query and the document. The coefficient c_1 used in the current version of the Cheshire II system is 1.269.

$$X_2 = \sqrt{QL}$$

This is square root of the query length (i.e., the number of terms in the query disregarding stopwords). The c_2 coefficient used is -0.310.

$$X_3 = \frac{1}{M}\sum_{j=1}^{M} \log DAF_{t_j}$$

This is the log of the absolute frequency of occurrence for term t_j in the document averaged over the M common terms. The c_3 coefficient used is 0.679.

$$X_4 = \sqrt{DL}$$

This is square root of the document length. In Cheshire II the raw size of the document in bytes is used for the document length. The c_4 coefficient used is -0.0674.

$$X_5 = \frac{1}{M}\sum_{j=1}^{M} \log IDF_{t_j}$$

This is the log of the *inverse document frequency* (IDF) for term t_j in the document averaged over the M common terms. The IDF is calculated as the total number of documents in the database, divided by the number of documents that contain term t_j. The c_5 coefficient used is 0.223.

$$X_6 = \log M$$

This is the log of the number of common terms. The c_6 coefficient used in Cheshire II is 2.01.

These coefficients and elements of the ranking algorithm have proven to be quite robust and useful across a broad range of document types.

Probabilistic searching, as noted above, requires only a natural language statement of the searcher's topic, and thus no formal query language or Boolean logic is needed for such searches. However, 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. At present, combined probabilistic and Boolean search results are evaluated using the assumption that the Boolean retrieved set has an estimated P(R | Q_{bool},D) = 1.0 for each document in the set, and 0 for the rest of the collection. The final estimate for the probability of relevance used for ranking the results of a search combining Boolean and probabilistic strategies is simply:

$$P(R \mid Q, D) = P(R \mid Q_{bool}, D) \cdot P(R \mid Q_{prob}, D)$$

where $P(R \mid Q_{prob},D)\$$ is the probability estimate from the probabilistic portion of the search, and $P(R \mid Q_{bool},D)$ the estimate from the Boolean. This has the effect of restricting the results to those items that match the Boolean portion, with ordering based on the probabilistic portion.

Relevance feedback is 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. Similarly, multiple records may be selected and submitted for feedback searching. In this

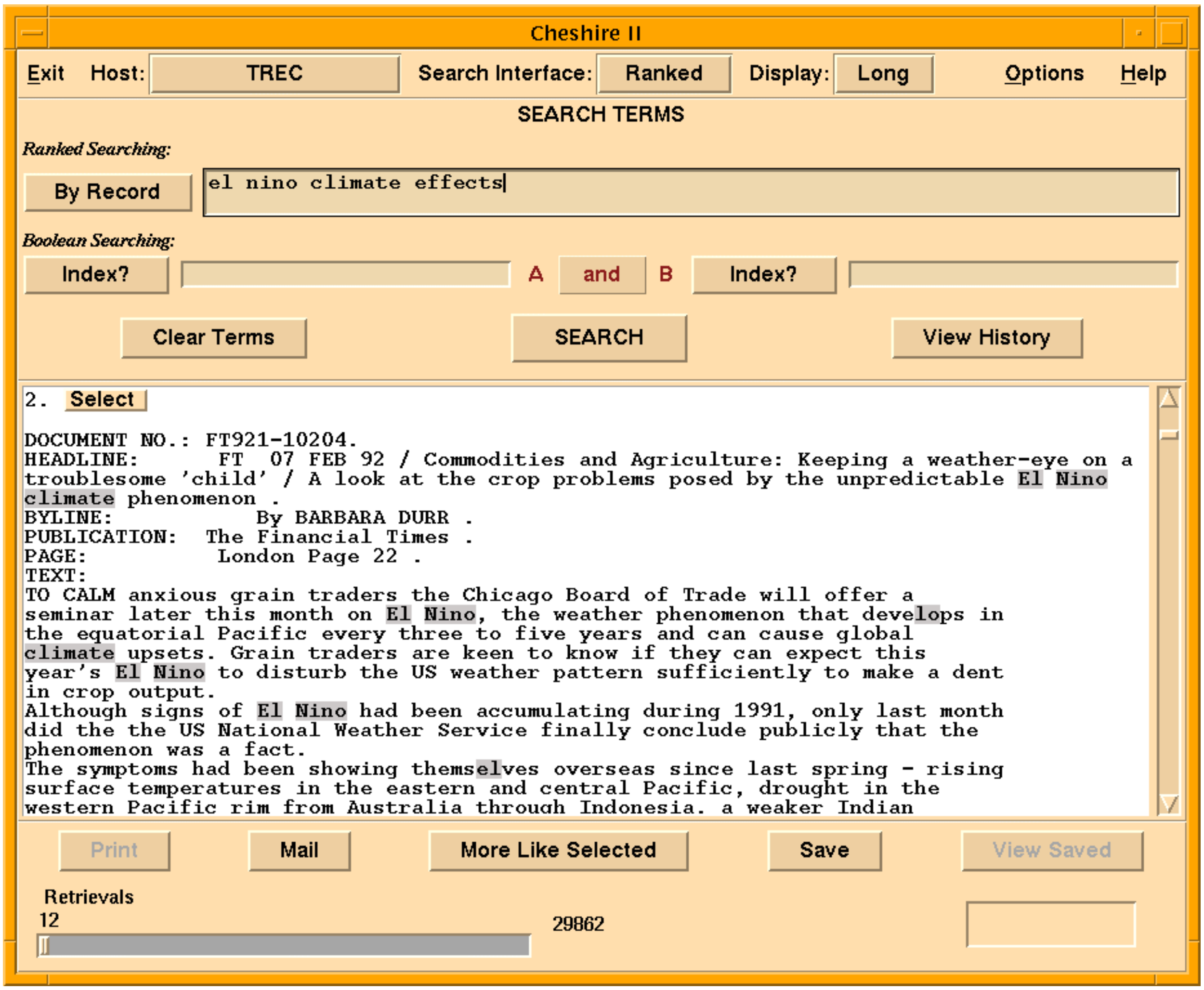

Figure 2: Cheshire II Long Display

case the contents of all those records are merged into a single query and submitted for searching. A feedback search is accomplished by parsing the selected record(s) and extracting the record elements specified for the index used for topical searching (as specified in the database configuration file). Each of these record elements is combined to form a single query, which is then submitted to the same probabilistic retrieval process described above. 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), but we plan to experiment further with various enhancements to our relevance feedback method.

The Cheshire II Client Interface

The design of the Cheshire II client interface (shown with the TREC FT database in Figures1and 2, was driven by a number of goals:

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;
4. to provide functions not immediately related to searching, such as print and e-mail facilities, to facilitate users' ability to 'take the results home'; and
5. to design a help system within the interface that would assist users not only in the mechanics of operating the Cheshire II client, but also in the more general tasks of selecting appropriate resources for searching, formulating appropriate queries, and employing various search tactics.

However, the initial design goals for the client interface made the assumption that most of the information that would be 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 TREC-7 Interactive track experiments was added to the existing interface easily, but as the comments and questionnaire responses discussed below show, this was probably not an optimal implementation for the tasks posed by the experiments.

Additional functionality beyond searching and browsing has been relatively easy to implement. Functions for printing, e-mailing and saving records are all available when records are displayed, and the user has the option of acting on either the entirety of the current record display or a subset thereof by selecting individual records using the "select" buttons on each record (visible in Fig. 1 next to the record numbers).

The Cheshire II client interface has been primarily implemented using the interpreted Tcl/Tk language (Ousterhout 1994), with a variety of lower-level functions, including the majority of the Z39.50 client interactions, written in the C programming language. This combination has proven quite successful in both providing the ability to rapidly prototype and modify the graphic user interface to accommodate new features, and to maintain a relatively high level of performance for the Z39.50 client-server interactions. Among the changes made to the client interface for TREC was the inclusion of display formats for the FT records were designed (as shown in Fig. 1 and Fig. 2). A routine was also added to highlight query terms in the text of the document to aid searchers in scanning for relevant passages. Note that the highlighting feature doesn't necessarily catch all of the terms that contributed to the selection of the document, because only the original query terms, and not stemmed terms, are used in the highlighting. Since the highlighting is using simple string matching on the text, partial words are sometimes highlighted erroneously.

The Zprise System

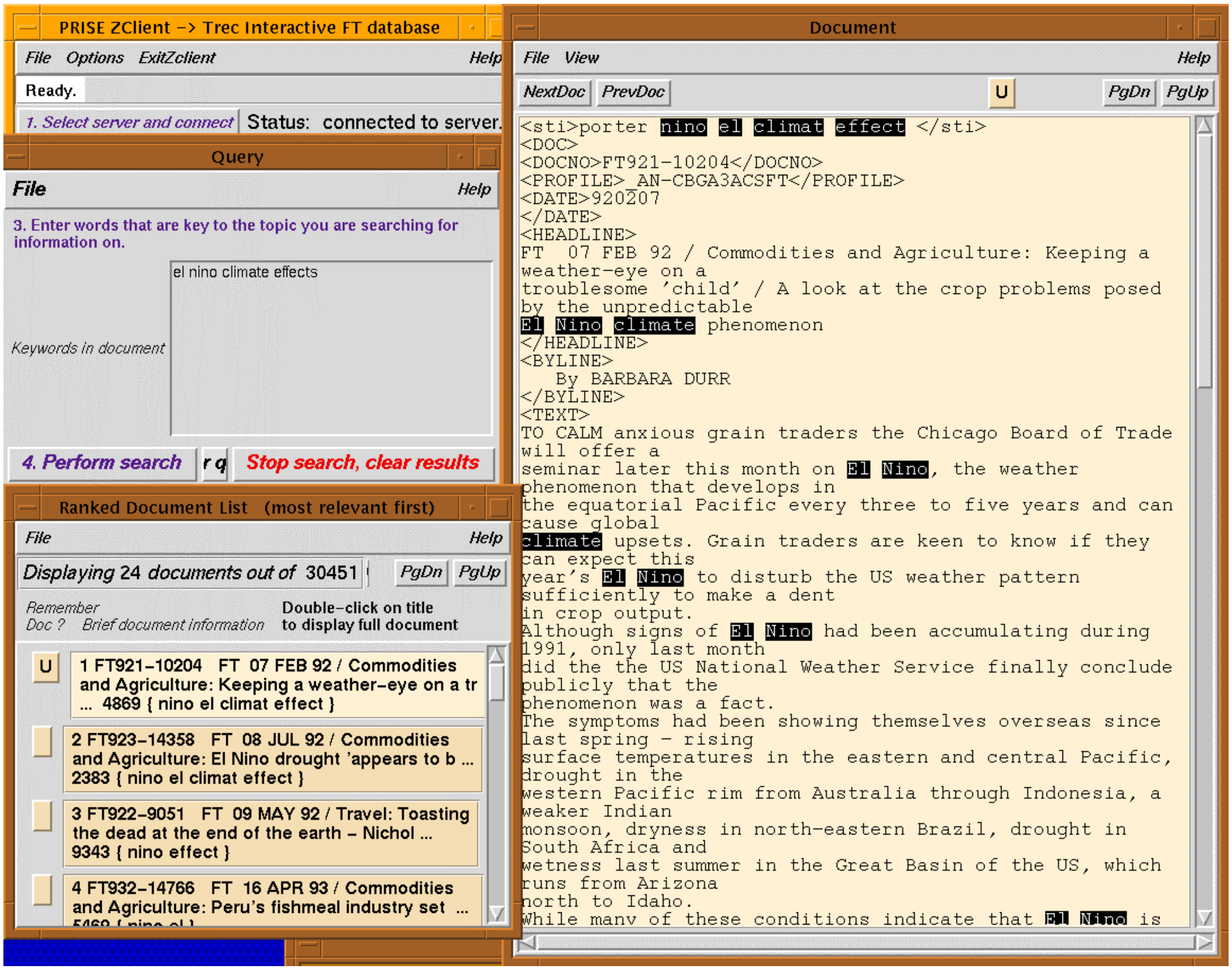

Figure 3. ZPRISE Display

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 the TREC-6 Interactive Track. 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 operation 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, while the Cheshire II interface, as noted above, was designed for brief metadata records and not with the idea of providing support for the sort of reading and selection activities that make up the user tasks in the TREC Interactive Track. In some ways the comparison between the interfaces comes down to how well a generic interface, not particularly adapted to the specific task, compares to an interface tailored to that task. The underlying PRISE search engine in ZPRISE uses (apparently) a fairly standard Vector Space model search algorithm, which performs quite well given the usually brief and simple query statements that characterized most searches by the searchers participating in this year's interactive track. In the following section I will describe the results obtained from the Aspectual Precision and Recall evaluations at NIST and the results of the demographic, search and system related questionnaires filled out by the participating searchers.

TREC Interactive Track

The administration of the interactive track followed the protocols set down in the track guidelines. This mandated a minimum group of 8 participant searchers, each of whom conduct 8 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. Because of some delays in obtaining the license and materials for the FA-1 Controlled Associations Test, this test was administered to the subjects independently from their actual search sessions.

The pooled results for all systems were evaluated at NIST by the TREC evaluators and "Aspectual

Asp. Precision	Topic								
System	352i	353i	357i	362i	365i	366i	387i	392i	Grand Total
a	0.4418	0.3860	0.4870	0.5913	0.9168	0.9168	0.8125	0.6110	0.6454
b	0.7448	0.3483	0.4125	0.7500	0.9018	0.8875	0.7375	0.8310	0.7017
C	**0.7993**	**0.2145**	**0.6368**	**0.7058**	**0.9793**	**0.9375**	**0.8458**	**0.8068**	**0.7407**
clus	0.4355	0.3594	0.5514	0.6773	0.8750	0.4291	0.5938	0.6276	0.5686
irisa	0.7448	0.4333	0.5715	0.7500	0.7223	1.0000	0.9063	0.8088	0.7421
irisp	0.7333	0.1750	0.5533	0.6368	1.0000	0.9168	0.7500	0.7085	0.6842
iriss	0.6745	0.6021	0.5803	0.6194	0.8875	0.8611	0.7884	0.7303	0.7179
J24	0.6250	0.5178	0.5083	0.5568	0.9375	0.8750	0.6615	0.7315	0.6767
list	0.2811	0.1916	0.4200	0.5209	0.9584	0.3750	0.8084	0.7540	0.5387
MB	0.8831	0.4876	0.4164	0.7249	0.7524	0.9679	0.6494	0.6020	0.6855
MR	0.6342	0.4309	0.2857	0.7056	1.0000	1.0000	0.7944	0.7190	0.6962
ok_noRF	0.9018	0.4984	0.3033	0.4981	0.8840	0.8750	0.7710	0.4874	0.6524
ok_withRF	0.8578	0.5865	0.3818	0.3558	0.7520	1.0000	0.8335	0.7475	0.6893
RUINQ-G	0.5844	0.5109	0.3618	0.4628	0.8814	0.7918	0.8854	0.6228	0.6357
RUINQ-R	0.5558	0.3160	0.4391	0.6055	0.8674	0.7889	0.7689	0.6819	0.6286
Z	**0.9500**	**0.4405**	**0.5030**	**0.6043**	**1.0000**	**0.8333**	**0.9333**	**0.7093**	**0.7467**
ZP	0.5863	0.4045	0.4285	0.8928	0.9688	0.9500	0.8873	0.8493	0.7459
zp_noRF	0.8875	0.2323	0.4303	0.7640	0.8750	0.7918	0.9015	0.8250	0.7134
Grand Total	0.6733	0.4090	0.4367	0.6324	0.8928	0.8354	0.7776	0.6916	0.6689

Table 1: Aspectual Precision for all TREC-7 Interactive Track Participants

Precision" and "Aspectual Recall" for each searcher was calculated. Table 1 shows the values for Aspectual

Precision by TREC topic for all systems in the TREC Interactive Track. Table 2 shows the values for

Aspectual Recall	Topic								
System	352i	353i	357i	362i	365i	366i	387i	392i	Grand Total
a	0.1250	0.2728	0.3658	0.3540	0.8023	0.5355	0.4723	0.3403	0.4085
b	0.2413	0.1593	0.3080	0.2918	0.8543	0.5000	0.3888	0.2433	0.3733
C	**0.2768**	**0.1820**	**0.3080**	**0.3123**	**0.8543**	**0.5358**	**0.4443**	**0.4373**	**0.4188**
clus	0.0893	0.1024	0.2310	0.1459	0.6874	0.2144	0.1110	0.1840	0.2207
irisa	0.2858	0.1593	0.2888	0.2708	0.5208	0.2860	0.1943	0.2430	0.2811
irisp	0.2413	0.1138	0.3655	0.1873	0.7918	0.5358	0.2220	0.3403	0.3497
iriss	0.1653	0.2161	0.2791	0.2605	0.7186	0.3571	0.3193	0.2985	0.3268
J24	0.1875	0.2275	0.3080	0.1875	0.8750	0.2503	0.4445	0.5280	0.3760
list	0.0534	0.0683	0.2791	0.1354	0.6770	0.0715	0.2499	0.2846	0.2274
MB	0.3368	0.1950	0.2090	0.3274	0.6161	0.4490	0.4522	0.2401	0.3532
MR	0.3061	0.1949	0.1155	0.2737	0.7260	0.3776	0.3171	0.4603	0.3464
ok_noRF	0.3349	0.2161	0.2406	0.3020	0.8489	0.4288	0.4305	0.2675	0.3837
ok_withRF	0.4105	0.2275	0.3655	0.3123	0.8230	0.3930	0.3610	0.2848	0.3972
RUINQ-G	0.2619	0.2340	0.2909	0.2499	0.8288	0.3930	0.4723	0.2918	0.3817
RUINQ-R	0.2365	0.2427	0.2214	0.3333	0.7449	0.3651	0.4073	0.2560	0.3489
Z	**0.2233**	**0.2955**	**0.2695**	**0.2918**	**0.8230**	**0.4285**	**0.6113**	**0.4723**	**0.4269**
ZP	0.2768	0.1593	0.3465	0.1668	0.8543	0.4643	0.5555	0.4308	0.4068
zp_noRF	0.3838	0.1138	0.3080	0.1878	0.8333	0.3930	0.5000	0.4723	0.3990
Grand Total	0.2478	0.1875	0.2573	0.2592	0.7503	0.3749	0.3750	0.3238	0.3472

Table 2. Aspectual Recall for all TREC-7 Interactive Track Participants

Aspectual Recall for all of the participating systems. Note that these two tables were derived from the per-search Recall and Precision figures reported by NIST. Note also that in these tables all of the system usages were combined in the calculations, therefore “ok_noRF” which was used in two separate experiments has the results from both experiments combined. 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 marginally better than the experimental system in terms of the Aspectual Precision. It is also interesting to note that virtually identical performance was achieved by the “ZP” system from NMSU and the “zp_noRF” from the Okapi Group, I believe that both of these systems, like “Z”, are unmodified ZPRISE systems.

The Cheshire II system also did not perform as well the control ZPRISE system in these experiments. This fact can largely be attributed to the more complex interactions required to perform the search tasks on the generic Cheshire II interface than on the ZPRISE system. In addition, there were some specific search failures due to misspelling (one searcher had 0 Precision and Recall for one search due to this).

Table 3 shows the mean and standard deviation of precision and recall over all searches for each searcher and system. As can be seen from the table, there was considerable range of performance within the searchers at Berkeley. In the following section we will examine the characteristics of the searchers as reported in the questionnaires administered during the experiments. Figure 4 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 8 participants. While none of the participants had used either the experimental (Cheshire II) or control (ZPRISE) systems in searching tasks, many 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 8 volunteers were scheduled for search sessions. A pre-search questionnaire asked each participant about:

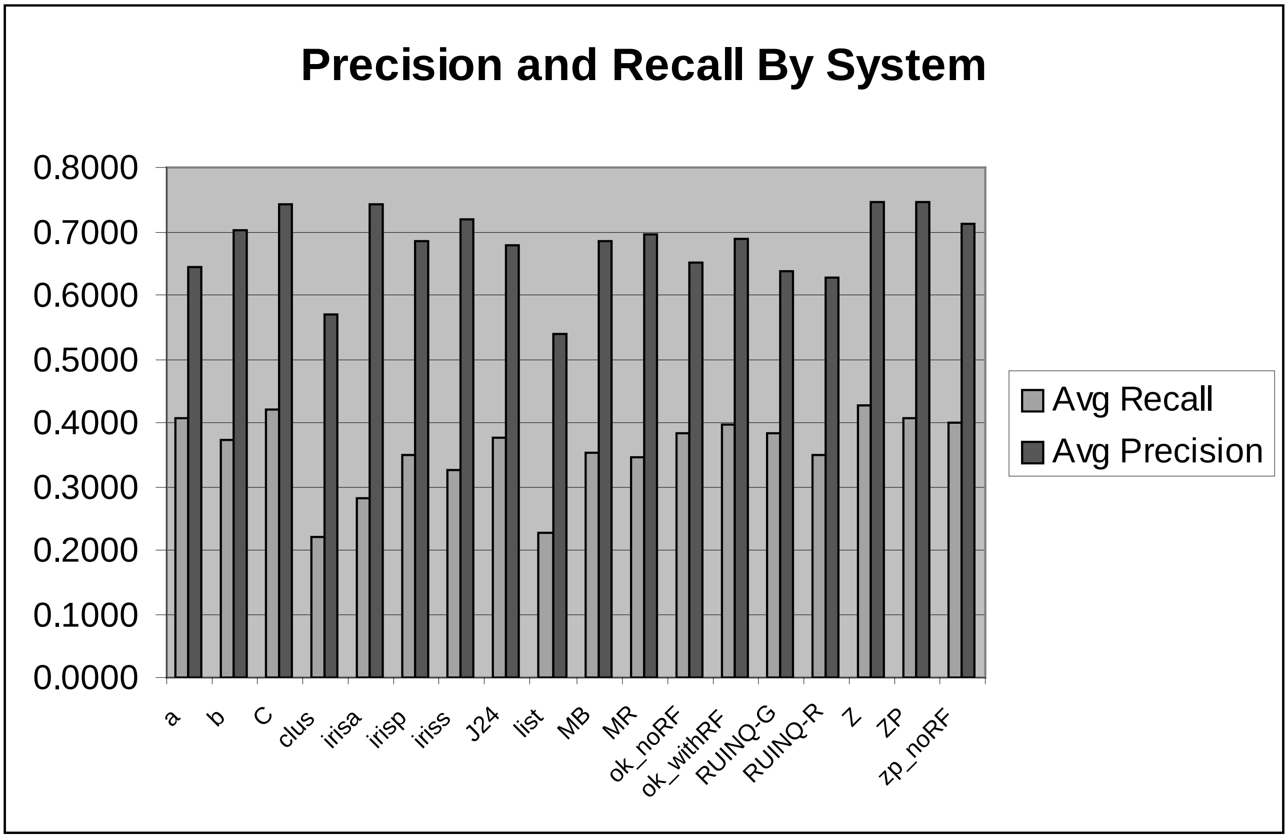

Figure 4. Average Precision and Recall by System

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, two others were PhD students with previous undergraduate and graduate degrees, and the remaining 4 were Masters students in the SIMS program). Two of the participants (P1 and P3) had considerable experience in online searching on other systems, the other two had very limited experience with online systems. The range of search experience with various systems is shown in Figure 6. By far, the most frequently used search systems were the Web search services and the next most frequent were online catalogs. Exploratory correlation analyses were performed on all of the variables from the presearch questionnaires, combined with the matching search and system questionnaires with the per-search and search precision and recall information from the NIST evaluators. Not surprisingly, there were very few significant correlations found in the analysis and many of those were trivial (years of search experience is positively correlated with age). Somewhat more interesting was that search experience with online systems like Dialog and BRS was also significantly correlated with search experience. 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.

The responses from each participant are shown in Appendix A. Table 4 shows the average responses for the "easy to do the search" question by searcher and system. As may be seen from the table, all 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. Once again, the ZPRISE system is given higher marks than the Cheshire system. Table 6 shows the average responses to the question "Are you

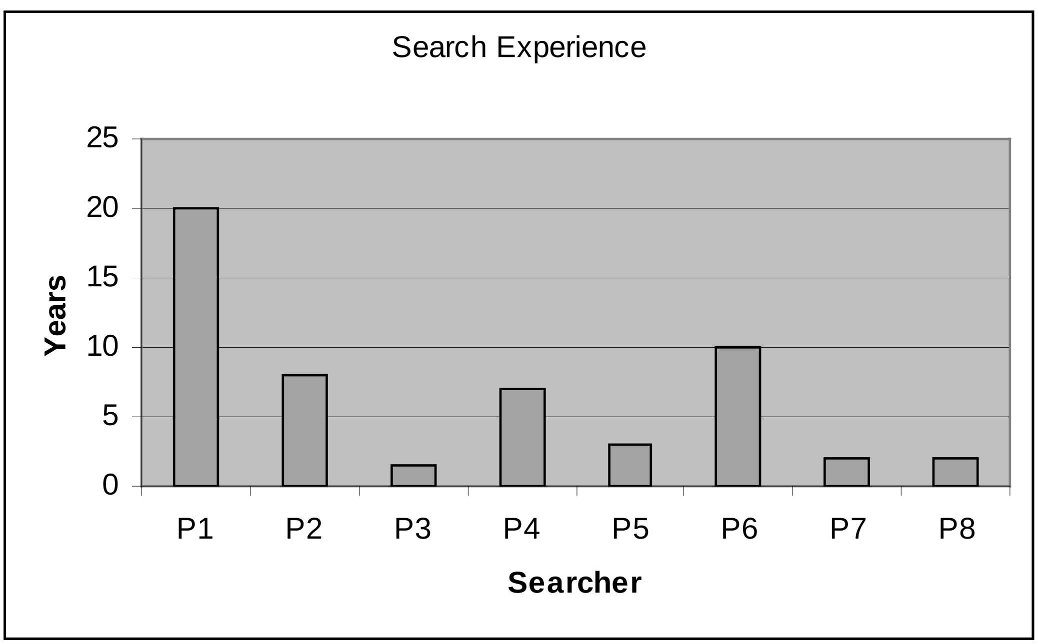

Figure 5. Search Experience

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. Satisfaction was however fairly strongly correlated with how easy it was to do the search task (Pearson's R=0.646, prob=0.0001). Interestingly, there was no significant correlation between any of the post-search questions and either Aspectual Precision or Aspectual Recall, but the signs of these correlations indicated some interesting items for further research. For example, a slight negative correlation was indicated between Precision and Recall and the user's confidence that they had identified all of the different instances for a topic.

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 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 system 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 missing features (e.g. Boolean AND but no NOT), and several searchers mentioned the need to scroll back to the beginning of a record to select it as relevant. Others (who used ZPRISE first) mentioned a preference for having the full-text document in a separate window. The following table shows the mean responses to these questions.

Question	C	Z	Mean
Mean of Easy to learn	4.00	4.50	4.25
Mean of Easy to use	3.38	4.25	3.81
Mean of Understand how to use	4.00	4.00	4.00

Mean Responses to Post-System Questionnaires

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 4.25), and they found the task similar to other searching tasks (mean of 3.63). They also found the systems somewhat different (mean of 3.37). In ranking the systems, 5 out of 8 ranked ZPRISE as easier to learn to use, while 7 out of 8 chose ranked it as easier to use. Curiously the searchers were evenly split (four each) on which they liked the best. One search commented that she would prefer Cheshire “for serious research” but found ZPRISE better suited to the TREC search tasks.

		System		
Searcher	Data	C	Z	Grand Total
P1	Precision Mean	0.7363	0.7338	0.7350
	Precision StdDev	0.1237	0.2347	
	Recall Mean	0.5443	0.5978	0.5710
	Recall StdDev	0.3205	0.1477	
P2	Precision Mean	0.8265	0.8375	0.8320
	Precision StdDev	0.2154	0.1974	
	Recall Mean	0.4488	0.4143	0.4315
	Recall StdDev	0.2190	0.1246	
P3	Precision Mean	0.7833	0.8125	0.7979
	Precision StdDev	0.3146	0.3750	
	Recall Mean	0.3420	0.3423	0.3421
	Recall StdDev	0.0645	0.3011	
P4	Precision Mean	0.7043	0.6875	0.6959
	Precision StdDev	0.2473	0.3750	
	Recall Mean	0.3568	0.3748	0.3658
	Recall StdDev	0.0664	0.3367	
P5	Precision Mean	0.6833	0.6368	0.6600
	Precision StdDev	0.2849	0.4404	
	Recall Mean	0.3345	0.4345	0.3845
	Recall StdDev	0.2842	0.2976	
P6	Precision Mean	0.8750	0.7083	0.7916
	Precision StdDev	0.1443	0.2765	0.2228
	Recall Mean	0.4238	0.3610	0.3924
	Recall StdDev	0.3330	0.1876	0.2525
P7	Precision Mean	0.6713	0.7740	0.7226
	Precision StdDev	0.3811	0.2785	0.3138
	Recall Mean	0.5565	0.4400	0.4983
	Recall StdDev	0.3460	0.2412	0.2831
P8	Precision Mean	0.6458	0.7833	0.7145
	Precision StdDev	0.4429	0.3146	0.3632
	Recall Mean	0.3440	0.4505	0.3973
	Recall StdDev	0.2492	0.2661	0.2454
Total Precision Mean		0.7407	0.7467	0.7437
Total Precision StdDev		0.2651	0.2893	0.2753
Total Recall Mean		0.4188	0.4269	0.4228
Total Recall StdDev		0.2430	0.2314	0.2354

Table 3: Average Precision and Recall by Searcher and system

Conclusions

Naturally, it impossible to draw any firm general conclusions from our small sample of searchers and searches. But it is obvious that an interface that is well adapted to the specific search task will tend to be

preferred by searchers even if the underlying system produces better overall performance in terms of Recall, and comparable performance in terms of precision. As observed at TREC-6, the overall performance of the Cheshire II system was quite good, although it was not dramatically better than the control system on average. These results, as has often been noted in previous TREC interactive evaluations, tend to be highly influenced by individual behavior and search techniques (this is apparent in the differences between the searchers on the same questions and in the same systems). What seems apparent from the results of the questionnaires coupled with the Precision and Recall measures is that a generic interface can perform quite acceptably in the TREC tasks, even if it isn't particularly liked by the users, compared to another system that is better suited to the TREC tasks, but may not be as useful in other situations. Since the task of reading and extracting relevant elements from texts is obviously important in a number of situations, I plan to add (among other interface suggestions from the users) a separate browsing window for full text documents.
.

Acknowledgements

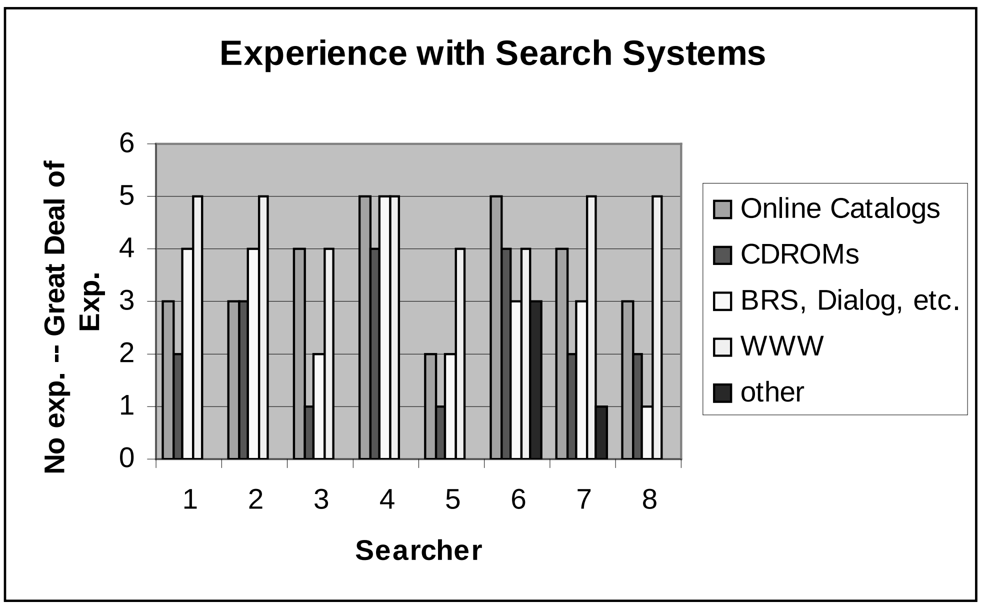

Figure 6. Experience with Search Systems

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.

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.

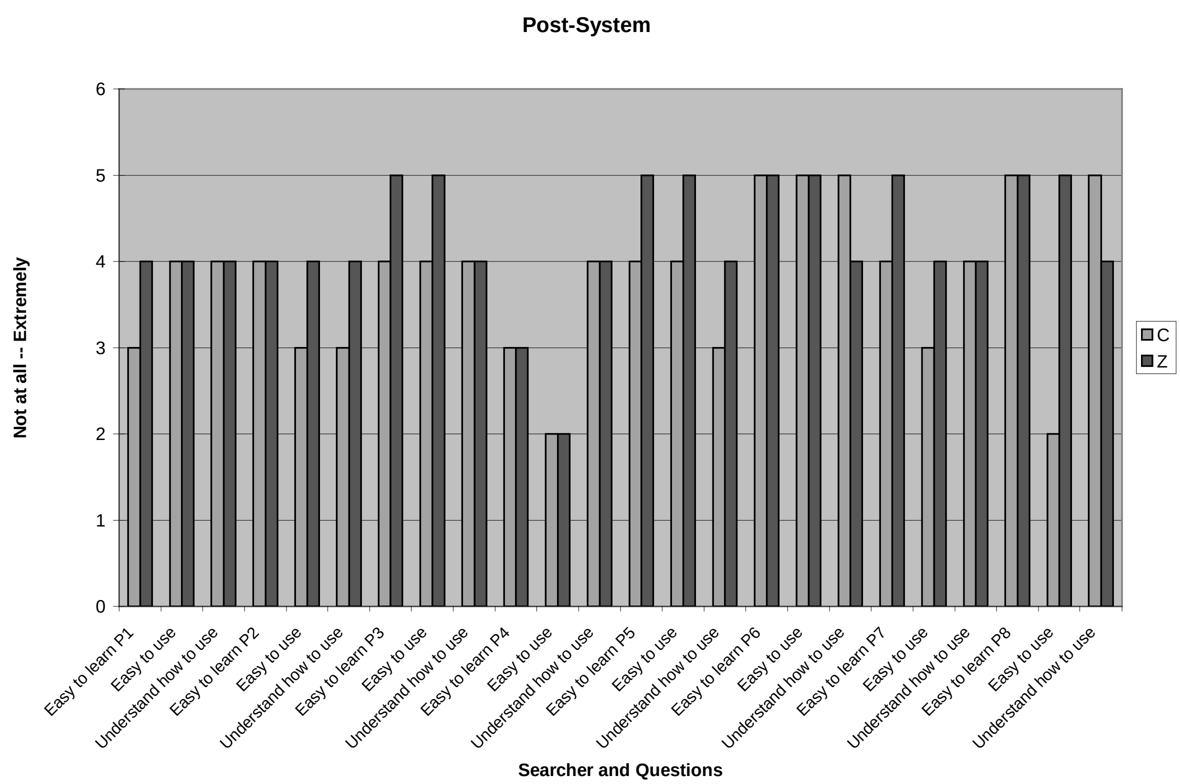

Figure 7: Post-system questions by searcher

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.*

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.

Average of easy_search	System		
Searcher	C	Z	Grand Total
P1	2.75	3.50	3.13
P2	2.75	3.00	2.88
P3	4.00	4.50	4.25
P4	3.25	3.50	3.38
P5	2.75	4.00	3.38
P6	2.00	3.75	2.88
P7	3.50	4.00	3.75
P8	2.75	4.50	3.63
Grand Total	2.97	3.84	3.41

Table 4: Was it easy to do the search on this topic?

Average of satisfied	Syste m		
Searcher	C	Z	Grand Total
P1	2.50	2.50	2.50
P2	2.25	3.25	2.75
P3	2.75	3.25	3.00
P4	2.25	2.50	2.38
P5	2.50	3.25	2.88
P6	1.75	2.00	1.88
P7	3.00	3.75	3.38
P8	2.75	4.00	3.38
Grand Total	2.47	3.06	2.77

Table 5: Are you satisfied with your search results

Average of familiar	System		
Searcher	C	Z	Grand Total
P1	2.00	2.25	2.13
P2	2.25	2.25	2.25
P3	1.75	2.50	2.13
P4	2.75	2.50	2.63
P5	2.50	2.75	2.63
P6	2.75	2.25	2.50
P7	1.50	3.25	2.38
P8	3.00	3.50	3.25
Grand Total	2.31	2.66	2.48

Table 6: Are you familiar with this topic

APPENDIX A: Responses to Post-Search Questionnaires by Searcher and Topic

		Topic								
Searcher	Data	352i	353i	357i	362i	365i	366i	387i	392i	Grand Total
P1	Familiar with Topic	2	3	1	2	3	1	2	3	2.13
	Easy to get started	3	4	3	3	4	3	3	3	3.25
	Easy to search	3	4	3	2	3	4	3	3	3.13
	Satisfied with results	2	2	2	2	4	3	2	3	2.50
	Confident of finding all	1	1	1	1	1	1	1	1	1.00
	Had enough time	4	3	4	3	4	4	3	3	3.50
P2	Familiar with Topic	2	2	2	2	3	2	3	2	2.25
	Easy to get started	3	2	3	3	3	2	4	3	2.88
	Easy to search	2	3	3	4	2	2	4	3	2.88
	Satisfied with results	1	3	3	3	2	3	4	3	2.75
	Confident of finding all	1	3	3	3	2	1	4	3	2.50
	Had enough time	2	2	3	2	2	1	4	3	2.38
P3	Familiar with Topic	3	1	1	3	3	1	2	3	2.13
	Easy to get started	5	5	4	5	5	3	5	5	4.63
	Easy to search	4	5	4	5	5	3	4	4	4.25
	Satisfied with results	3	3	3	4	3	2	3	3	3.00
	Confident of finding all	3	2	3	4	2	1	1	1	2.13
	Had enough time	3	3	4	5	5	1	3	1	3.13
P4	Familiar with Topic	2	2	1	4	3	1	4	4	2.63
	Easy to get started	4	4	4	4	3	4	4	4	3.88
	Easy to search	4	4	3	4	3	3	3	3	3.38
	Satisfied with results	2	3	3	2	3	2	2	2	2.38
	Confident of finding all	2	3	2	2	1	2	2	2	2.00
	Had enough time	2	3	2	2	4	2	2	2	2.38
P5	Familiar with Topic	3	3	1	3	3	3	4	1	2.63
	Easy to get started	2	4	3	3	2	4	4	4	3.25
	Easy to search	2	4	3	3	3	4	4	4	3.38
	Satisfied with results	3	2	3	3	1	3	4	4	2.88
	Confident of finding all	1	2	2	2	1	5	4	3	2.50
	Had enough time	3	5	3	2	2	5	5	4	3.63
P6	Familiar with Topic	1	4	4	2	4	1	3	1	2.50
	Easy to get started	1	2	3	5	3	4	3	5	3.25
	Easy to search	2	2	2	1	3	4	4	5	2.88
	Satisfied with results	2	1	2	1	2	2	1	4	1.88
	Confident of finding all	1	1	1	1	1	3	1	1	1.25
	Had enough time	1	1	1	1	1	1	1	1	1.00
P7	Familiar with Topic	3	2	3	4	3	1	2	1	2.38
	Easy to get started	5	5	4	5	5	3	4	4	4.38
	Easy to search	2	4	4	5	5	3	4	3	3.75
	Satisfied with results	2	4	4	4	5	3	3	2	3.38
	Confident of finding all	2	4	4	4	5	2	3	1	3.13
	Had enough time	3	4	3	4	5	2	1	1	2.88
P8	Familiar with Topic	4	1	4	4	2	4	3	4	3.25
	Easy to get started	4	5	5	4	4	3	5	4	4.25
	Easy to search	4	1	5	4	5	3	3	4	3.63
	Satisfied with results	4	1	5	3	4	3	3	4	3.38
	Confident of finding all	2	4	1	1	1	3	3	3	2.25
	Had enough time	2	5	1	1	5	3	4	3	3.00
Total Familiar with Topic		2.50	2.25	2.13	3.00	3.00	1.75	2.88	2.38	2.48

Total Easy to get started	3.38	3.88	3.63	4.00	3.63	3.25	4.00	4.00	3.72
Total Easy to search	2.88	3.38	3.38	3.50	3.63	3.25	3.63	3.63	3.41
Total Satisfied with results	2.38	2.38	3.13	2.75	3.00	2.63	2.75	3.13	2.77
Total Confident of finding all	1.63	2.50	2.13	2.25	1.75	2.25	2.38	1.88	2.09
Total Had enough time	2.50	3.25	2.63	2.50	3.50	2.38	2.88	2.25	2.73