Published paper
Joho, H., Sanderson, M. and Beaulieu, M. (2004) *A study of user interaction with a concept-based interactive query expansion support tool.* In: Advances in Information Retrieval : 26th European Conference on IR Research, ECIR 2004, Sunderland, UK, April 5-7, 2004. Proceedings. Lecture Notes in Computer Science (2997). Springer , Berlin / Heidelberg, pp. 42-56.

A Study of User Interaction with a Concept-based Interactive Query Expansion Support Tool

Hideo Joho, Mark Sanderson, and Micheline Beaulieu

Department of Information Studies, University of Sheffield,
Regent Court, 211 Portobello Street, Sheffield, S1 4DP.
{H.Joho|M.Sanderson|M.Beaulieu}@sheffield.ac.uk

Abstract. A medium-scale user study was carried out to investigate the usability of a concept-based query expansion support tool. The tool was fully integrated into the interface of an IR system, and designed to support the user by offering automatically generated concept hierarchies. Two types of hierarchies were compared with a baseline. Several observations were made as a result of the study: 1) the hierarchy is often accessed after an examination of the first page of search results; 2) accessing the hierarchies reduces the number of iterations and paging actions; 3) accessing the hierarchies increases the chance of finding relevant items more accurately than the baseline; 4) the hierarchical structure helps the users to handle a large number of concepts; and finally, 5) subjects were not aware of the difference between two types of hierarchies.

1 Introduction

In interactive query expansion (IQE), users often find it difficult to select expansion terms from a suggested list [1, 2]. Possible reasons for this is that the statistical weighting tends to generate low frequency, specific, or unfamiliar terms, and the list does not provide the context for the suggested terms. However, our previous study and others suggest that the hierarchical organisation of candidate expansion terms can offer better both context and greater efficiency in the query expansion process [3, 4]. This paper presents a user study of a concept-based approach to IQE.

CiQuest (Concept-based Interactive QUery Expansion Support Tool) is a support system for interactive searches. It provides an overview of a set of retrieved documents which allows the user to focus on a particular subset of the search results. It also provides a set of candidate terms that can be used to replace or expand a user's initial query. The CiQuest system is designed to achieve these two facilities through concept hierarchies. A concept hierarchy is *dynamically* generated from a set of retrieved documents and visualised by cascading menus. More general terms are placed at a higher level followed by related but more specific terms at a lower level.

Our overall research aim is to study the use of a concept-based system to support information retrieval. The specific objectives are to:

- evaluate the retrieval effectiveness of document derived concept structures for selecting relevant documents in a retrieved document set;

- evaluate the retrieval effectiveness of incorporating concept structures to assist users in selecting candidate terms for interactive query expansion; and
- assess how searchers make use of concept structures to bridge the gap between the query space and the document space in interactive searching.

The next section will discuss our experimental methodology including the details of our system and experimental design. The results and analysis of our experiments will then be presented. The paper concludes with an overall discussion of our findings and future work.

2 Experimental Design

The Interactive Track of TREC (Text REtrieval Conference[1]) has been developing a test collection for research into interactive information retrieval. We used the test collection from TREC-8 Interactive Track [5] as well as the Ad-hoc task as the basis of our experiments. It consists of six topics, relevance information, and a collection of 210,158 articles (564MB of texts) from the Financial Times 1991-1994. Each topic contains a title, description, and definition of instances as shown in Fig. 1.

The task defined by the TREC-8 Interactive Track is referred to as an *instance finding task*. In this task the subjects are asked to find as many different instances or answers to the query as possible, as opposed to finding as many relevant documents as possible as in the Ad-hoc task. For example, Topic 408i is designed to find the instances of the tropical storms that have caused property damage or loss of life. The subjects are also asked to save at least one document for each of the different aspects or answers of the topic.

Fig. 1. TREC-8 Interactive Track Sample Topic (408i)

Number: 408i
Topic: tropical storms
Description: What tropical storms (hurricanes and typhoon) have caused property damage and/or loss of life?
Instances: In the time alloted, please find as many DIFFERENT storms of the sort described above as you can. Please save at least one document for EACH such DIFFERENT storm. If one document discusses several such storms, then you need not save other documents that repeat those, since your goal is to identify as many DIFFERENT storms of the sort described above as possible.

2.1 Participants

Twelve participants were recruited from Department of Information Studies and Computer Science, and included two females and ten males who were either research stu-

[1] http://trec.nist.gov/

dents or research assistants. Their educational qualification included one with a PhD, eight with a Master, and three with a Bachelor. Of the twelve, two had participated a TREC experiment before but neither had experience of seeing the topics and tasks used in our experiment.

2.2 System and Interface Development

The CiQuest system is a tool designed to support information access through two basic functionalities: multi-document summarisation and interactive query expansion. Words and noun phrases (i.e. *concepts*) are extracted from the retrieved documents and used to form a hierarchical structure which, as a whole, can be seen as a summary of search results. Individual concepts that are organised in a *general to specific* manner and can also be seen as candidate terms to expand or reformulate initial queries.

The core technology of the system is to determine the semantic specificity of concepts with little human involvement or knowledge resources. Our overall aim is to find a pair of related concepts and determine which is more general (or specific). A hierarchy is, thus, formed as a result of the cumulation of such a process. For our experiment we have implemented two different approaches for the generation of the hierarchies.

Generating hierarchies The first approach is based on the statistical analysis of document frequency and co-occurrence information between concepts, and called the subsumption approach which was originally developed by Sanderson and Croft [6]. In this approach, concept C_i is said to subsume concept C_j when a set of documents in which C_j occurs is a subset of the documents in which C_i occurs, or more specifically, when the following two conditions are held: $P(C_j \mid C_i) \geq 0.8$[2] and $P(C_i \mid C_j) < 1$.

The assumption is that C_i is likely to be more general than C_j because, first, the former appears more frequently than the latter, and second, the former subsumes a large part of C_j's document set. Also they are likely to be related since they co-occur frequently within documents. A similar assumption has been made by other researchers (e.g. [7, 8]. A sample hierarchy using this approach can be found in Fig. 2.

The second approach is called the trigger phrase approach, and is based on the lexical and syntactic analysis of noun phrases which have been found to be useful for query expansion [9]. A trigger phrase is a phrase that matches a fragment of text that contains a parent-child description. Words and phrases found in the description are used to formulate the hierarchy. Our trigger phrases are based on Hearst [10] who originally used them to find additional lexical relations in WordNet [11]. Examples of the phrase patterns are:

- SUCH AS: ... international organisations **such as** WHO, NATO, and ...
- AND OTHER: ... WHO, NATO, **and other** international organisations are ...
- INCLUDING: ... international organisations, **including** WHO, NATO, and ...

In the above example, when one of the patterns is matched, the concept *international organisations* is set as a superordinate of *WHO* and *NATO* in the above example.

[2] This value was set by them empirically.

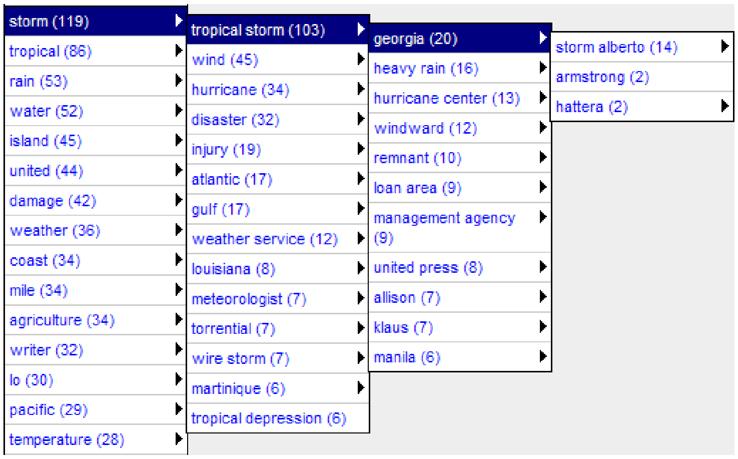


Fig. 2. Sample hierarchy generated by the subsumption approach with the top 200 documents retrieved in response to the query *tropical storm*. The number next to the term indicates the frequency of occurrence. You can see the phrase ”tropical storm” is subsumed by the term ”storm”. Also several instances of storms or hurricanes such as *george*, *allison*, or *klaus* are successfully organised under ”tropical storm”.

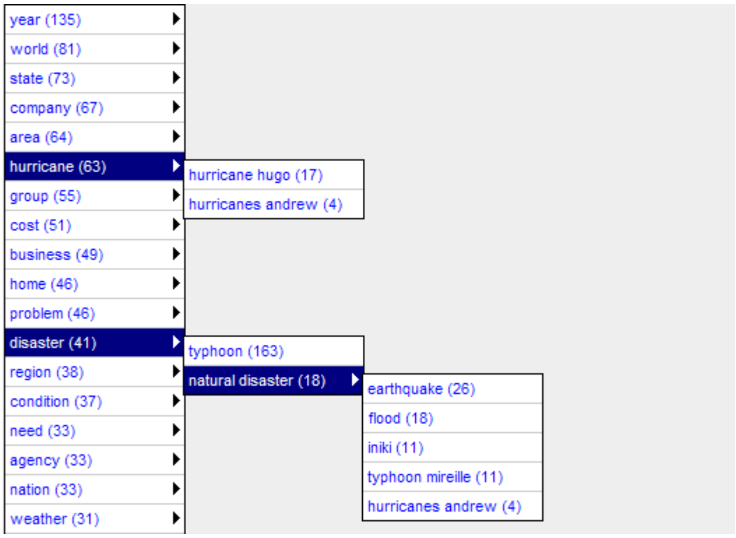
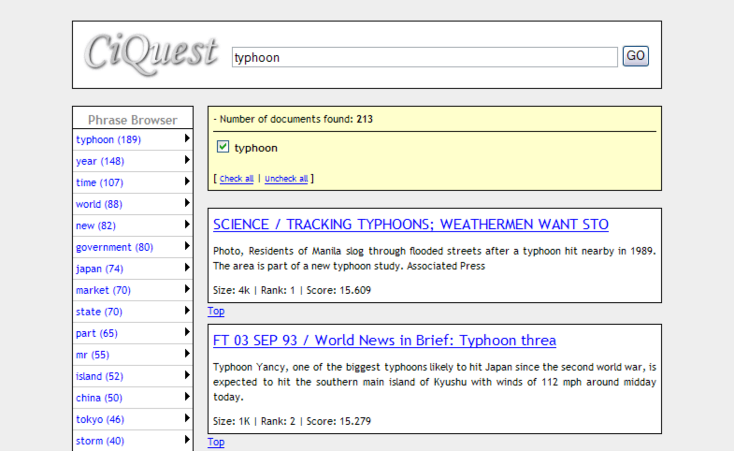


Fig. 3. Sample hierarchy generated by the trigger phrase approach with the top 200 documents retrieved in response to the query *typhoon hurricane*. Noun phrases such as *hurricane hugo* and *hurricane andrew* can be found under the head noun ”hurricane” at the top level of the hierarchy. Also you can find the terms such as *earthquake*, *flood*, and phrases including *typhoon* or *hurricane* organised as an instance of ”natural disaster”.

Furthermore, the head noun of phrases is identified and set as a superordinate of the phrases (similar to [12]). For example, *organisations* (head noun) is set as a superordinate of *international organisations*. This head noun extraction also helps the hierarchy to include more phrases that contains the same head noun. In other words, this approach attempts to generate a hierarchy of noun phrases using the lexical evidence and the head nouns. A sample hierarchy using this approach can be found in Fig. 3.

CiQuest system in use Once a hierarchical structure of related concepts is generated, the system visualises it using cascading menus. The top level of hierarchies are shown in the left side of the main result page (See Fig 4). Our principle regarding the integration of the hierarchy into an IR system's interface is to provide the functionality without disturbing the default search process. The default search process is to submit a query, look through the hitlist, and open a page to access the fulltext.



- Number of documents found: 213

☑ typhoon

SCIENCE / TRACKING TYPHOONS; WEATHERMEN WANT STO

Photo, Residents of Manila slog through flooded streets after a typhoon hit nearby in 1989. The area is part of a new typhoon study. Associated Press

Size: 4k | Rank: 1 | Score: 15.609

FT 03 SEP 93 / World News in Brief: Typhoon threa

Typhoon Yancy, one of the biggest typhoons likely to hit Japan since the second world war, is expected to hit the southern main island of Kyushu with winds of 112 mph around midday today.

Size: 1K | Rank: 2 | Score: 15.279

Fig. 4. CiQuest system: Top level of menu is shown along with the search result

Backend IR system: CiQuest system in the current paper was integrated into the Okapi system [13]. The best passage identified by the weighting scheme was displayed in every record of search results.

Browsing the hierarchy: When a mouse pointer is *placed* on a concept in the menu, a list of its subordinate concepts is displayed. The presence of subordinates is indicated by a small triangle arrow at the right-side of each entry.

Focusing on a subset: When a concept in the menu is *clicked*, a set of documents in which the concept occurs within the retrieved documents is shown in the same format as in the initial results. This subset of documents is also ordered by the ranking of the initial results. In this *focusing mode*, a pointer link is displayed at the bottom of the page to allow the user to go back to the initial results.

Refreshing the hierarchy: When another query is submitted, the hierarchy is automatically refreshed based on a set of documents retrieved in response to the new query.

2.3 Experimental Procedures

Experiments were based on the CiQuest system, but three different versions were devised for the test. The first was a baseline system which offered no support function. The second and third versions each incorporated the subsumption and the trigger phrase approaches respectively. Although the underlying functionality was different, subjects were not made aware of this as they searched through a common web-based interface.

Each test subject undertook searches on three TREC-8 topics, one to test each version of the system. The allocation of topics and test system was done randomly so that each topic was, thus, searched by six subjects. Participants were briefed on two tasks: the first was the *instance finding* tasks as described above. The second task, *query optimising*, required searchers to generate a so-called optimal or best query based on their search experience of the topic. The optimising task made it possible to compare the effectiveness of the optimal query with that of the initial query based on precision and recall for document relevance as used for the TREC Ad-hoc task, as opposed to the instance relevance used in the Interactive task.

The first exeriment, therefore, is the true interactive searching task, and the second experiment is a black-box input/output approach which does not take account of user interaction.

After the demonstration of the system, subjects were given several minutes to use the system with a sample topic. The subjects were then given 10 minutes for the instance finding task, but were allowed to take as long as they wish for the query optimising task. However, they tended to complete the task within a couple of minutes. Subjects also completed questionnaires at the beginning of the test session, after each search, and on completing the whole experiment. The questionnaires were based on the instruments developed for the TREC Interactive Track. The procedure took 60 to 90 minutes in total for each subject.

3 Results and Analysis

The results and analysis of our experiments using the precision/recall measures[3], log analysis, questionnaire, and manual observation are as follows. Three groups of the system settings as described above will be referred to as the *Baseline*, *Subsump* menu, and *Trigger* menu in this section.

3.1 Instance finding task

Instance recall and precision The instance recall is calculated based on the number of instances correctly identified by the subjects divided by the total number of instances

[3] Overall, it was rare to find the statistical significance using t-test due to the sample size, but it is indicated by a star (*) where applicable.

identified by the NIST assessors (called official instances). The instance precision is calculated based on the number of correctly identified instances divided by the total number of instances identified by the subjects.

Table 1. Instance recall and precision

		Baseline		Subsump		Trigger	
	Official	Instance	Instance	Instance	Instance	Instance	Instance
Topic ID	instances	recall	precision	recall	precision	recall	precision
408i	24	0.313	0.834	0.250	0.659	0.084	0.667
414i	12	0.375	0.729	0.292	0.875	0.459	0.745
428i	26	0.423	0.816	0.289	0.917	0.231	0.709
431i	40	0.138	0.625	0.113	0.625	0.175	0.399
438i	56	0.215	0.690	0.188	0.857	0.161	0.988
446i	16	0.188	0.715	0.282	0.700	0.313	0.410
Average		0.275	0.735	0.235	0.772	0.237	0.653

Table 1 shows the instance recall and precision of the three groups. Each topic was used by two subjects in all groups. Although the difference among the groups are generally small, the result shows that the Baseline's recall is higher than the menu groups while the Subsump achieved the highest precision among them.

As for the higher recall with the Baseline, two reasons can be possible. One is that the Okapi back-end IR system performed well [14], thus, the subjects could find relevant instances without support. Another is that the Baseline group could spent more time to examine a greater number of documents while the menu groups were spending the time browsing the hierarchies. However, the higher precision with the Subsump suggests that the accuracy of identifying relevant instances can be improved by the hierarchies.

Document access rate The subjects were asked to save a document in which they found one or more instances. Table 2 shows the number of documents in which subjects selected and viewed the full-text (called seen documents) and documents that were actually saved as relevant.

As can be seen, the subjects viewed more documents in the Baseline than the Subsump or Trigger but saved less frequently. With the menus the seen documents were more often saved. Here, with the previous table's result, we can see a trend of improving the accuracy of identifying relevant documents and instances when the hierarchies were used.

Interaction, paging, and access to the menu Table 3 shows the data about the iteration of searches, paging, and access to the menus, which provides additional insight of user behaviour in the instance finding task. An iteration is defined as a new query or refomulated query in the course of session. A paging is defined as moving one result page to another. A menu access is defined as clicking on a concept term to display the set of linked documents.

Table 2. Document access rate (%)

	Base			Subsump			Trigger		
Topic ID	Seen doc	Saved doc	Rate	Seen doc	Saved doc	Rate	Seen doc	Saved doc	Rate
408i	18.5	9.0	52.58	16.5	8.0	47.98	10.5	2.5	26.44
414i	11.0	4.5	40.00	6.0	2.5	41.43	10.5	3.5	36.12
428i	14.0	11.0	80.75	13.5	9.0	66.49	9.0	7.0	77.78
431i	17.5	8.5	49.02	11.5	7.5	67.50	8.0	4.5	73.08
438i	23.5	17.0	72.64	12.0	11.0	92.86	11.0	9.0	83.04
446i	13.5	5.0	38.93	12.5	6.0	47.73	12.0	10.5	87.77
Average	16.3	9.17	55.65	12.0	7.3	60.66	10.2	6.2	64.04

Table 3. Iteration, paging, and access to the menu

	Base		Subsump				Trigger			
Topic ID	Iter	Paging	Iter	Paging	Menu	Saved	Iter	Paging	Menu	Saved
408i	6.5	7.0	3.5	2.0	8.0	2.5	8.5	0.0	3.5	0.0
414i	5.5	5.5	2.5	2.0	4.5	0.5	3.0	1.5	4.0	1.0
428i	3.0	5.0	2.0	1.5	4.5	2.5	1.5	2.0	1.5	0.5
431i	4.5	2.0	4.0	1.5	0.5	0.0	4.0	2.5	3.0	1.0
438i	4.5	2.0	4.0	1.5	0.5	0.0	4.0	2.5	3.0	1.0
446i	3.0	4.5	3.0	1.5	6.0	0.5	3.5	4.0	1.0	0.0
Average	4.50	4.33	3.17	1.67	4.00	1.00	4.08	2.08	2.67	0.58

First, the number of iterations shows that the subjects submitted fewer queries with the menu groups than the Baseline. Also, the frequency of going to the next page in the Baseline is higher than the menu groups. Both, along with Menu access information, indicate that the menus were used to focus on a subset of documents as opposed to submitting a new query or going to the next pages. Saved access is the number of accesses to the menus which lead to save any documents (i.e. find an instance). In this regard, it appears that the Subsump performed marginally better than the Trigger menu.

Summary Overall, the results from the instance finding task suggests that the menus can be useful to accurately identity relevant information from search results, and reduce the number of iterations and paging actions (i.e. takes less effort).

3.2 Query optimising task

The query optimising task was evaluated using the relevance judgements of the TREC-8 Ad-hoc task. The purpose of this task was to compare the effectiveness of the optimal query with that of the initial query based on precision and recall for the full retrieved document sets, as opposed to the documents viewed and judged by the subjects.

Overall Table 4 shows the retrieval effectiveness of initial queries, which are the first query submitted by the subjects, and optimised queries, which the subject generated after searching each topic. This result confirmed that the subjects could improve their initial queries after 10 minutes of search experience.

Table 4. Overall performance of query optimisation

	Initial	Optimised	Diff.(%)
No. of session	36	36	
No . of Retrieved Rel docs	2512	3050	21.42*
Precision			
At 1 docs	0.5278	0.6111	15.80
At 5 docs	0.5333	0.5611	5.20
At 10 docs	0.4472	0.5056	13.00
At 20 docs	0.4069	0.4569	12.30
At 30 docs	0.362	0.3981	10.00
Avg. Prec	0.2029	0.2348	15.72

* indicates statistical significance at $p < 0.05$

Out of 36 sessions, 32 initial queries were modified and four were unchanged. Out of 32 changed queries, 20 had an increase of terms, 6 had a decrease, and 6 had no difference in number. The number of increased terms varies between one and three with the average of 1.45 terms. Although the overall changes against the initial queries were small, As can be seen in Table 4, these small changes contributed to the retrieval of a significantly larger number of relevant documents.

Table 5. Query optimisation across the systems

	Baseline			Subsump			Trigger		
	Initial	Opt.	Diff.(%)	Initial	Opt.	Diff.(%)	Initial	Opt.	Diff.(%)
No. of session	12	12		12	12		12	12	
Retrieved Rel	781	979	25.35	923	1033	11.92	808	1038	28.47
Precision									
At 1 docs	0.500	0.583	16.70	0.583	0.667	14.30	0.500	0.583	16.70
At 5 docs	0.517	0.550	6.50	0.517	0.533	3.20	0.567	0.600	5.90
At 10 docs	0.450	0.450	0.00	0.425	0.492	15.70	0.467	0.575	23.20
At 20 docs	0.396	0.425	7.40	0.400	0.454	13.50	0.425	0.492	15.70
At 30 docs	0.347	0.361	4.00	0.381	0.408	7.30	0.358	0.425	18.60
Avg. Prec.	0.195	0.224	14.97	0.208	0.228	10.01	0.206	0.252	22.17

Across the system setting Table 5 shows the comparison of initial and optimised queries over the three system settings. As expected the performance of initial queries were found to be similar across the systems and they were lower than the optimised queries. However, based on the previous task, we did not expect the Trigger menue session to outperform others. From the average precision we can see the Trigger menu contributed most in generating a better query, followed by the Baseline, and Subsump.

Across the topics Table 6 shows the retrieval effectiveness of both types of queries over six topics used in our experiment. Overall, the optimised queries outperformed the initial ones in all topics with the exception of Topic 408i.

An interesting point is that the improvement achieved by the optimised queries seems to be reasonably consistent across topics which had varied performances of the initial queries (e.g. from 0.1070 to 0.3383 in Average Precision). Although more data would be required to draw any conclusive comments, it seems that the optimised queries could improve the retrieval effectiveness regardless of the performance of initial results.

Table 6. Query optimisation across the topic

Topic	408i			414i			428i		
	Initial	Opt.	Diff (%)	Initial	Opt.	Diff (%)	Initial	Opt.	Diff (%)
No. of session	6	6		6	6		6	6	
Retrieved Rel	379	320	-15.57	212	188	-11.32	525	614	16.95
Precision									
At 5 docs	0.333	0.167	-50.00	0.500	0.567	13.30	0.633	0.700	10.50
At 10 docs	0.250	0.167	-33.30	0.367	0.500	36.40*	0.533	0.633	18.80
At 20 docs	0.317	0.167	-47.40	0.333	0.417	25.00	0.508	0.600	18.00
At 30 docs	0.339	0.183	-45.90	0.317	0.356	12.30	0.406	0.494	21.90
Avg Prec	0.147	0.088	-14.97	0.237	0.253	6.70	0.291	0.338	16.35
Topic	431i			438i			446i		
	Initial	Opt.	Diff (%)	Initial	Opt.	Diff (%)	Initial	Opt.	Diff (%)
No. of session	6	6		6	6		6	6	
Retrieved Rel	535	778	45.42	540	691	27.96	362	459	26.8
Precision									
At 5 docs	0.733	0.733	0.00	0.400	0.567	41.70	0.600	0.633	5.60
At 10 docs	0.717	0.683	-4.70	0.250	0.517	106.70	0.550	0.533	-3.00
At 20 docs	0.608	0.617	1.40	0.233	0.450	92.90	0.450	0.492	9.30
At 30 docs	0.550	0.533	-3.00	0.217	0.400	84.60	0.367	0.422	15.20
Avg Prec	0.338	0.418	23.59	0.109	0.182	66.32*	0.107	0.129	20.45

* indicates the statistical significance at $p < 0.05$.

Summary The results from the query optimising task shows that the learning curve for optimising their initial queries are similar among the three groups. However it appears that the Trigger group performed marginally better than the other two groups. The strongest trend of the improvements in the menu groups was found in the precision at the document level of 1 to 30 (in Table 5) while the Baseline group was likely to improve at the lower document levels.

This suggests two points. One is that the optimised queries generated by the menu groups could be based on the selection of the relevant documents from a wider range of rankings than the Baseline. Another possibility is that such optimised queries should stand a better chance to bring up the rankings of a wider range of relevant documents.

3.3 User perception

Now that the results based on the recall/precision and log analysis have been discussed, following two sections will present the results from the questionnaires and manual observations.

Subjects were asked to fill in a short questionnaire after each session. The following aspects of the CiQuest system were investigated by the questionnaire:

1. Ease of use of the system
2. Size of menus (Too long or too many?)
3. The menus as a tool to help predicting the contents of linked documents
4. The menus as a tool to help relevance judgement of documents
5. The menus as a tool to help focusing on important terms
6. The menus as a tool to help understanding the contents of documents
7. The menus as a tool to help having a better idea of a set of retrieved documents
8. Preference of system settings

The result of Question 1 to 7 is shown in Table 7.

Table 7. User perception (Score 1: Not at all, 4: Sometimes, 7: Always)

		Score										Score							
Question	Type	1	2	3	4	5	6	7	Average	Question	Type	1	2	3	4	5	6	7	Average
1	Subsump	1	0	1	2	7	0	1	4.50	5	Subsump	1	1	0	3	2	3	2	4.75
	Trigger	1	2	1	4	1	2	1	4.00		Trigger	2	0	3	2	1	3	1	4.08
2	Subsump	3	1	4	4	0	0	0	2.75	6	Subsump	2	1	1	2	4	2	0	3.92
	Trigger	1	4	1	3	2	1	0	3.33		Trigger	3	1	4	3	0	1	0	2.92
3	Subsump	1	0	0	4	4	3	0	4.58	7	Subsump	1	1	0	3	3	4	0	4.50
	Trigger	2	0	2	3	2	2	1	4.08		Trigger	2	1	3	2	1	3	0	3.67
4	Subsump	1	0	1	4	4	2	0	4.33										
	Trigger	2	0	3	2	2	2	1	4.00										

Use of system Question 1 asked the subjects how easy it was to use the system, rated between 1 (Not at all) and 7 (Always). The table shows that the Trigger menu's score is distributed across the scale, whereas the majority scored the Subsump menu at 5.

Size of menu Question 2 sought to establish to what extent the menus were considered to be too long or containing too many terms. The lower score is better in this question. The Subsump menu's score concentrated at the lower end of scales while the Trigger menu's ratings were distributed more widely. Nevertheless the size of the menus did not seem to overwhelm the subjects in either case.

Predicting contents Question 3 asked how useful a menu was to predict the contents of documents linked to the terms in a menu. The menu was designed to show a set of documents linked to each term in the menu when a user clicked it. As can be seen, the majority of subjects (11) gave a score between 4 and 7 for the Subsump menu. Although there were fewer subjects (8) for the Trigger menu who gave a score in this range, it appears that both types of menus succeeded in predicting the contents of linked documents.

Relevance judgement Question 4 asked how useful a menu was for judging the relevance of documents during the sessions. Although the instance finding task was not to find a relevant document, the task latently involved the assessment of relevance (i.e. no instance would be found in a non-relevant document). The table shows that more subjects with the Subsump menu gave a score between 4 and 7 than with the Trigger menu.

Focusing on important terms Question 5 asked how useful a menus was for focusing on terms of interest. As described before, the hierarchy provided a means of narrowing down to a subset of retrieved documents regardless of its ranked position. The scores of both types of menus were well distributed in the range above 4. The Subsump menu seemed to gain a slightly higher overall score than the Trigger menu.

Understanding contents Question 6 asked how useful a menu was to understand the contents of documents. The table shows that the scores for the Subsump menu are generally high with the score 5 as the peak while the Trigger menu has the peak at the score 3.

Better idea of retrieved documents Question 7 asked if a menu provided a better idea of a set of retrieved documents as opposed to individual documents. Similar to the previous question the Subsump menu seemed to gain a higher overall score than the Trigger menu.

Preference of system setting After the completion of all sessions the subjects were asked their preference among the three settings with the overall feedback against the system. Two points became clear from the final questionnaire. First, more than half of the subjects showed their preference for the Baseline system because of its simplicity and familiarity. Second, most subjects except two did not clearly notice the difference between the two types of menus in terms of how to organise terms. This point will be discussed further in a later section.

Summary The subjective evaluation of the hierarchies was presented through the questionnaires. Generally the subjects find the Subsump menu more useful than the Trigger menu in supporting information access. The scores of the Trigger menu tend to be distributed across the scale, while for the Subsump menu they are concentrated at a higher level. A more detail comparison of the concepts generated in the two hierarchies should be carried out to gain a better insight of how users interpret those concepts.

3.4 Other user behaviour

In addition to the precision/recall evaluation, analysis of system logs, and questionnaires, observations were made and recorded manually during the sessions, the following describe some typical user behaviours.

Accessing the hierarchies The most commong approach for accessing the hierarchy was:

1. Submit a query;
2. Examine several records in the first page of the results; then
3. Browse the hierarchy.

This route seems to show that the primary concern in the search process is on the documents. However it was found that many subjects decided to browse the menus after the first-page examination, as opposed to going on to the next page. This also seems to be influenced by the results of the first-page examination. When a subject found a reasonable amount of relevant documents in the first page, they tended to go to the next page. The hierarchy seemed to be accessed more frequently when the subjects were less satisfied with the first page.

Using the hierarchies It was observed that there were two typical ways of using the hierarchy. One was to focus on a subset of documents. This was the most popular way to use the menus as described above. However, another way was to assess the potential usefulness of terms. In other words, some subjects selected a term, examined the title and best paragraph of the top linked documents, selected another term, examined the list, and repeated this process.

Browsing the hierarchies The top level terms of the menu seem to be very important for the subjects in using the hierarchy. In particular it was observed that the absence of query terms at the top level seemed to discourage browsing through the hierarchy. This happened more often in the Trigger menu than in the Subsump menu. Hence, the top level terms were regarded as a starting point.

Another observation is that the subjects tended to go back to the same parent term when one of its children was found to be useful, and try another child term.

A final comment is that users' browsing action (i.e. movement from one concept to another) tended to be carried out easily and speedly. Although the subjects commented that they were aware of some irrelevant concepts included in the hierarchies, they seemed to be capable of filtering out those concepts during their tasks.

4 Conclusion and future work

4.1 Conclusive discussion

We presented a user study to investigate the usability of the CiQuest system that was designed to support interactive searches. Our focus was on the task-based evaluation of the system as well as the standard precision/recall measures. From the instance finding task, it was found that the Baseline was also effective due to the good performance of our IR system, but the precision can be improved with the hierarchies. The query optimising task indicated that the hierarchies could help improve the precision at the higher document levels (i.e. 5 to 30) more significantly than the Baseline.

Questionnaires and manual observation revealed that the hierarchical structures can be easily used and be useful to support the information accessing process. Also several interesting user behaviours that can be characteristic in the use of concept hierarchies were identified and discussed. The main highlights of our findings are:

1. the hierarchy is often accessed after an examination of the first page of search results;
2. accessing the hierarchies reduces the number of iterations and paging actions;
3. accessing the hierarchies increases the chance of finding relevant items more accurately than the Baseline;
4. the hierarchical structure helps the users to handle a large number of concepts; and finally,
5. subjects were not aware of the difference between two types of hierarchies.

4.2 Future work

In the query optimising task the Trigger hierarchy seems to slightly outperform the Subsump hierarchy. However the questionnaire indicated that the Subsump hierarchy was preferable. This suggests both approaches have can be beneficial as a means of generating a concept hierarchy to support information retrieval. Therefore an integration of two approaches is worthy a further investigation.

Exploring other techniques to determine the hierarchical relations between concepts should also be examined. For example, we came across the research by Bookstein [15] during the development of our system. Their analysis of symmetric and asymmetric relations between terms by measuring clumping strength could also be of interest.

5 Acknowledgements

The authors thank to the participants of our experiments from Department of Information Studies and Computer Science at the University of Sheffield. This work was carried out as a part of the CiQuest project, which was funded by the Library and Information Commission (now *re*:source, the Council for Museums, Archives and Libraries). Any opinions, findings, and conclusions described here are the authors and do not necessarily reflect those of the sponsor.

References

1. Belkin, N.J., Cool, C., Head, J., Jeng, J., Kelly, D., Lin, S., Lobash, L., Park, S.Y., Savage-Knepshield, P., Sikora, C.: Relevance feeback *versus* local context analysis as term suggestion devices: Rutgers' trec-8 interactive track experience. In Voorheer, E.M., Harman, D.K., eds.: Proceedings of the 8th Text REtrieval Conference (TREC-8), Gaithersburg, MD, NIST (2000)

2. Ruthven, I.: Re-examining the potential effectiveness of interactive query expansion. In Callan, J., Cormack, G., Clarke, C., Hawking, D., Smeaton, A., eds.: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Tronto, Canada, ACM (2003) 213–220
3. Pollitt, S.: Interactive information retrieval based on faceted classification using views. In: Proceedings of the 6th International Study Conference on Classification (FID/CR), London, UK, University College of London (1997) Available from http://scom.hud.ac.uk/external/research/CeDAR/dorking.htm [Accessed: 08/01/2004].
4. Joho, H., Coverson, C., Sanderson, M., Beaulieu, M.: Hierarchical presentation of expansion terms. In: Proceedings of the 17th ACM Symposium on Applied Computing (SAC'02), Madrid, Spain, ACM (2002) 645–649
5. Hersh, W., Over, P.: Trec-8 interactive track report. In Voorheer, E.M., Harman, D.K., eds.: NIST Special Publication 500-246: The Eighth Text REtrieval Conference (TREC-8), Gaithersburg, ML, NIST (2000) 57–64
6. Sanderson, M., Croft, B.: Deriving concept hierarchies from text. In Hearst, M., Gey, G., Tong, R., eds.: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Berkeley, CA, ACM (1999) 206–213
7. Niwa, Y., Nishioka, S., Iwayama, M., Takano, A.: Topic graph generation for query navigation: Use of frequency classes for topic extraction. In: Proceedings of the Natural Language Processing Pacific Rim Symposium (NLPRS'97), Phuket, Thailand (1997) 95–100
8. Nanas, N., Uren, V., De Roeck, A.: Building and applying a concept hierarchy representation of a user profile. In Callan, J., Cormack, G., Clarke, C., Hawking, D., Smeaton, A., eds.: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Tronto, Canada, ACM (2003) 198–204
9. Anick, P.G., Tipirneni, S.: The paraphrase search assistant: Terminological feedback for iterative information seeking. In Hearst, M., Gey, G., Tong, R., eds.: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Berkeley, CA, ACM (1999) 153–161
10. Hearst, M.A.: Automatic acquisition of hyponyms from large text corpora. In: Proceedings of the Fourteenth International Conference on Computational Linguistics, Nantes, France (1992) 539–545
11. Miller, G.A.: Wordnet: A lexical database for english. Communications of the ACM **38** (1995) 39–41
12. Wacholder, N., Evans, D.K., Klavans, J.L.: Automatic identification and organization of index terms for interactive browsing. In: Proceedings of the first ACM/IEEE-CS joint conference on Digital libraries, Roanoke, VA, ACM (2001) 126–134
13. Robertson, S., Walker, S., Hancock-Beaulieu, M.: Large test collection experiments on an operational, interactive system: Okapi at trec. Information Processing & Management **31** (1995) 345–360
14. Fowkes, H., Beaulieu, M.: Interactive searching behaviour: Okapi experiment for trec 8. In Robertson, S., Ayse, G., eds.: Proceedings of the BCS-IRSG 22nd Annual Colloquium on Information Retrieval Research, Cambridge, UK, BSC-IRSG (2000) 47–56
15. Bookstein, A., Kulyukin, V., Raita, T., Nicholson, J.: Adapting measures of clumping strength to assess term-term similarity. Journal of the American Society for Information Science and Technology **54** (2003) 611–620