Research Report
Joho, H., Sanderson, M. and Beaulieu, M. (2003) *Concept-based Interactive Query Expansion Support Tool (CIQUEST).* Library and Information Commission Research Report 149. ISBN 1-902394-79-8.

Concept-based Interactive Query Expansion Support Tool (CIQUEST)

by

Micheline Beaulieu, Mark Sanderson, and Hideo Joho

Department of Information Studies
University of Sheffield

Library and Information Commission Research Report 149

RE/090

ISBN 1-902394-79-8
ISSN 1466-2949

This report is available from http://ciquest.shef.ac.uk/Licrr149.pdf.

Library and Information Commission Research Reports are published by Resource: The Council for Museums, Archives and Libraries and may be purchased as photocopies or microfiche from the British Thesis Service, British Library Document Supply Centre, Boston Spa, Wetherby, West Yorkshire LS23 7BQ, UK.

Abstract
This report describes a three-year project (2000-03) undertaken in the Information Studies Department at The University of Sheffield and funded by Resource, The Council for Museums, Archives and Libraries. The overall aim of the research was to provide user support for query formulation and reformulation in searching large-scale textual resources including those of the World Wide Web. More specifically the objectives were: to investigate and evaluate methods for the automatic generation and organisation of concepts derived from retrieved document sets, based on statistical methods for term weighting; and to conduct user-based evaluations on the understanding, presentation and retrieval effectiveness of concept structures in selecting candidate terms for interactive query expansion.

The TREC test collection formed the basis for the seven evaluative experiments conducted in the course of the project. These formed four distinct phases in the project plan. In the first phase, a series of experiments was conducted to investigate further techniques for concept derivation and hierarchical organisation and structure. The second phase was concerned with user-based validation of the concept structures. Results of phases 1 and 2 informed on the design of the test system and the user interface was developed in phase 3. The final phase entailed a user-based summative evaluation of the CiQuest system.

The main findings demonstrate that concept hierarchies can effectively be generated from sets of retrieved documents and displayed to searchers in a meaningful way. The approach provides the searcher with an overview of the contents of the retrieved documents, which in turn facilitates the viewing of documents and selection of the most relevant ones. Concept hierarchies are a good source of terms for query expansion and can improve precision. The extraction of descriptive phrases as an alternative source of terms was also effective. With respect to presentation, cascading menus were easy to browse for selecting terms and for viewing documents. In conclusion the project dissemination programme and future work are outlined.

Authors

Micheline Beaulieu is a Professor of Information Science and leads the Information Retrieval Research Group in the Department of Information Studies at the University of Sheffield. Prior to taking up her current post in 1998, she was co-director of the Interactive Systems Research Centre at City University for over ten years working principally on the evaluation and development of the Okapi probabilistic retrieval system. Both departments have been rated as 5* centres of excellence in IR research. Professor Beaulieu has published widely in the following areas: the design and evaluation of interactive information retrieval systems, information seeking behaviour, information retrieval and HCI and the application of IT in the Humanities.

Dr. Mark Sanderson is a senior lecturer within the Department of Information Studies at the University of Sheffield. He is co-editor of ACM SIGIR Forum, and is on the editorial board of ACM TOIS (Transactions On Information Systems), JASIST (Journal of the American Society of Information Science and Technology), IP&M (Information Processing and Management), and IR (Information Retrieval). Prior to his current post, he was a research assistant for four years, working first at the University of Glasgow and then at the Center for Intelligent Information Retrieval in the University of Massachusetts, USA. His PhD on Word Sense Disambiguation was carried out at the University of Glasgow. Since arriving at Sheffield, he joint coordinator of the IST 2000 project Clarity, developing Cross language information retrieval tools. He is also a principal investigator on the MIND, SPIRIT and Eurovision projects.

Mr. Hideo Joho is a research assistant at the Department of Information Studies at the University of Sheffield. He received an MSc in Information Management from the Department in 1999. He is currently involved in the SPIRIT project, developing a spatially-aware search engine. He is also doing his PhD in information retrieval.

Concept-based Interactive Query Expansion Support Tool (CIQUEST)

Project report to
Re:source The Council for Museums, Archives and Libraries

Micheline Beaulieu , Mark Sanderson, and Hideo Joho

Department of Information Studies
The University of Sheffield.
m.beaulieu/m.sanderson@sheffield.ac.uk

June 2003

Acknowledgment

The CiQuest research team is grateful to Re:Source and its predecessor the Information and Library Commission who awarded the grant to support this research.

CONTENTS

1. Introduction

This report describes a three-year project (2000-2003) undertaken in the Information Studies Department at The University of Sheffield and funded by Re:Source, the Council for Museums, Archives and Libraries. The Computational Informatics Research Group successfully responded to a call for proposals for research in Information Retrieval (IR) by the Library and Information Commission, the predecessor of Re:Source.

An ongoing debate in IR has been concerned with determining the relative merits of keyword versus concept-based retrieval approaches. Although the latter based on manually constructed thesauri or subject classification schemes have primarily served as indexing tools, more recently these have also been explored as search tools to support user query formulation (Kristensen 1993, Jones et al 1995). Both approaches are evident on the World Wide Web (WWW). Search engines such as AltaVista and Excite rely on basic Boolean keyword retrieval, although additional techniques such as term weighting and relevance ranking are also commonly applied as in Google for example. Retrieval in Yahoo however is primarily dependent on metadata or manually constructed subject hierarchies. The limitations of both approaches as experienced by Web searchers are well documented (Jansen et al 1998). Moreover the human assignment of subject categories inherently lack the specificity of concepts that searchers often require for effective retrieval. On the other hand work on generating thesauri automatically from document collections (Evans & Lefferty 1994, Grefenstette 1994) indicates that searchers are reluctant to explore such concept-based tools for query formulation.

An alternative line of research, which has proven to be more productive in supporting user query formulation, has been automatic or interactive query expansion (AQE, IQE) based on relevance feedback (Robertson et al 1997). However this approach also has limitations. Its success depends on three factors: the user identifying good relevant documents, the relevant documents containing appropriate terms to add to the query and in the case of IQE, the effective display of candidate terms for user selection.

The current project aimed to address some of these issues by introducing an innovative concept-based approach to query expansion. The project investigated and evaluated the automatic generation and hierarchical organisation of concepts derived from retrieved documents to supporting user query formulation and reformulation. The objective was to provide users with an overview of retrieved documents to assist them in finding relevant documents and to display potential query terms in a meaningful context.

The next sections of the report present the aims and objectives of the project, methodology adopted and a discussion of the issues addressed and outcomes of each of the four project phases. The concluding section assesses the impact of the work, dissemination plans and describes continuing research.

2. Aims and Objectives

Given that users generate broad and brief queries and then encounter difficulties in refining their initial query on the basis of items retrieved, the overall research aim was to provide user support for query formulation and reformulation in searching large-scale textual resources including those on the WWW. The specific objectives were:

- To investigate methods for the automatic generation and organisation of concepts derived from retrieved documents;
- To explore the presentation of document derived concept structures for incorporation into a user interface;
- To evaluate the retrieval effectiveness of document derived concept structures for selecting relevant documents in a retrieved document set;
- To evaluate the retrieval effectiveness of incorporating concept structures to assist users in selecting candidate terms for interactive query expansion;
- To assess how searchers make use of concept structures to bridge the gap between the query space and the document space in interactive searching

3. Methodology

The overall methodological approach was to address the interdependent research issues related to the five stated project objectives in a progressive and integrated fashion and to incorporate an evaluative component in each of the project phases. A special emphasis was put on user participation in the different elements of the investigation and to take full account of user searching behaviour and the interactive searching process as well as retrieval effectiveness testing in the approach to evaluation.

The TREC test collection formed the basis for different aspects of the various evaluative experiments conducted and included ad-hoc topics from TREC 6,7 and 8. In addition two large-scale retrieval tests were conducted using test data retrieved directly from the WWW by search engines including Google, WebTop and DirectHit. This was deemed to be a more realistic approach than using the TREC Web Track dataset. The user-based experiments adopted the general task orientated design approach established by the TREC Interactive Track.

Seven experiments were conducted in the course of the project encompassing four distinct phases, which addressed specific research issues related to the project aim and objectives. The research context, methodological approach and outcomes for each phase will be discussed below.

4. Experimental phases and results

The project work plan was divided into four phases over a period of 32 months. In the first phase a series of experiments was conducted to investigate further techniques for concept derivation and hierarchical organisation and structure. The second phase was concerned with user-based validation of the concept structures. Results of phases 1 and 2 informed on the design of the test system and user interface developed in phase 3. Phase 4 was a user-based summative evaluation of the CiQuest system.

4.1 Phase 1: Concept derivation and structure

As outlined above the project built on previous work on the creation of concept hierarchies using co-occurrence information (Sanderson and Croft 1999). Initial results from a user study had shown that approximately half of the hierarchy's concept relationships had been sensibly arranged. Having established that co-occurrence was a relatively simple but successful method for creating concept hierarchies, the next step was to examine other ways of increasing the number of meaningful relationships. Two aspects were investigated. The first aimed to improve concept derivation by using a simple pattern matching descriptive phrase finder for proper nouns. The second sought to validate the structure or organization of the concept relationships by comparing text pairs of related concepts (i.e. parent/child relationships) created by the simple statistical approach with those incorporated in the WordNet thesaurus. The underlying assumptions were that simple techniques are likely to be much faster than those that require natural language processing techniques such as parsing and that simple but accurate methods are likely to be most beneficial in dealing with very large corpora.

4.1.1 Descriptive phrase finder experiments

A descriptive phrase finder (DPF) system was developed and tested in a series of three experiments. The main feature of the system is that it uses simple pattern matching to identify key proper noun phrases. However a ranking algorithm based on accuracy, sentence position and frequency occurrence across documents enhances the pattern matching.

First experiment: TREC LA Times test collection

The first experiment focused on two aspects of the effectiveness of the DPF: the accuracy of the system output and an initial investigation of the impact of size of collection on accuracy. Although the TREC LA Times test collection was used as a data set, 50 queries were generated by the experimenters to ensure that the proper noun phrases covered the time frame of the documents. Relevance judgements were based on full sentences and not just on extracted descriptions. The results showed that the system was able to rank a description-bearing sentence within the top ten for 94% of the queries. However as expected the effectiveness of the system was reduced as the collection size got smaller in line with results obtained in the Trec VLC collection. The experiment was reported in:
Joho, H. and Sanderson. Retrieving descriptive phrases from large amounts of free text. In: *Agah, A. Callan, J & Rundenstenier, E. (eds). Proceeding of the 9th International Conference on Information and Knowledge Management, McLean,*

Virginia, 2000. New York: Association for Computing Machinery, 2000, 180-186. See Appendix 1.

Second experiment: Web using Google

Having established the feasibility of the simple approach for locating descriptive phrases from free-text, in the second experiment the approach was tested on the Web using the Google retrieval engine. The retrieved web documents were fetched from a maximum of 600 URLs returned by Google for each query. Two criteria for relevance were used. *Strict relevance* was based on clearly defined descriptive phrases and *lenient relevance* accepted variant phrases. For example the key answer for the query Tony Blair, `Tony Blair is the current Prime Minister of the United Kingdom' was regarded as strictly relevant but `Tony Blair is a political leader' would be a lenient relevance judgement.

The results for successfully answered queries in the top 5, 10 and 20 sentences using 50 queries based on strict relevance was 66%, 82% and 88% respectively. On the other hand the results using the lenient relevance definition was 100% for the top 5 and 10, a distinct improvement on the first experiment using the LA times clearly indicating that the size of the collection does impact on the effectiveness of the system. It would appear that there is a better chance of locating a relevant descriptive phrase in a larger collection and that cross document term occurrence statistics contribute to improving performance. The experiment was reported in: Joho, H., Liu, Y.K. and Sanderson, M. Large scale testing of a descriptive phrase finder. *In: Allen, J. (ed). Proceedings of the 1st Human Language Technology Conference, San Diego, California, 2001*. Morgan Kaufmann, 2001, 219-221. See Appendix 2.

Third experiment: Impact of search engine ranking and number of URLs

In the third experiment attention was focused on two aspects: firstly to what extent the Google's PageRank system impacted on the effectiveness of the DPF system, and secondly to what extent the different number of retrieved URLs could affect performance. The ranking techniques of DPF were further tested by examining how three search engines ranked a common set of web pages for each query. Intersection sets from Google, WebTop and DirectHit were created by retrieving 1000 URLs per engine, noting the common URLs and their rank position and extracting the top 55 ranked URLs from each engine. The same queries as in experiment 2 were used but because of the lack of intersections, only 28 were used for the analysis. Comparisons at 100, 210 and 600 URLs generally indicate that the percentage of successfully answered queries and the precision in the top 20 sentences improves with the increase in the number of URLs. With regard to the ranking method of the different search engines, WebTop and Google performed equally well at the top 20 sentences but Google achieved higher precision. It would thus appear that different search engine ranking methods does also affect the performance of the DPF system. However further analysis would be required on a larger overlapping collection. This experiment is reported in a draft paper, Joho, H, Liu, T. T., Liu, Y. K., Sanderson, M, and Beaulieu, M. Descriptive phrase finder: a free-text mining tool for descriptive information about proper nouns. To be submitted to the *Journal of the American Society for Information Science and Technology*. See Appendix 3.

4.1.2 Concept relationships experiments

The second part of phase 1 compared the simple statistical approach for creating parent/child relationships with the linguistic approach of the WordNet thesaurus. In particular we examined the relationship between document frequency and term specificity for creating topic hierarchies. The question to be addressed was, given a pair of term/concepts that have been found to be related, how does one determine which is the more specific?

The aim was to measure on a large scale the accuracy of document frequency for ordering word pairs taken from WordNet. Some 45,000 noun words and phrases in WordNet were submitted as queries in Google. Document frequency for a synset was then estimated by averaging the number of documents retrieved for each member term. In the experimental design, it was assumed that the commonest sense of a term accounts for the great majority of that term's occurrences and that this would correspond to the definition of commonest sense in WordNet. Among the semantic relations in WordNet, hypernymy (superordinate) orders words by their specificity. Thus disambiguated hypernym chains were used as the basic data for three heuristic experiments.

First experiment: Average document frequency

The average document frequency of synsets was calculated for each level of a set of hypernym chains that were the same length. Although it was expected that there would be a monotonically decreasing line between the most general level and the most specific level, it was found that the middle level had the second highest average. It is not entirely clear why that should be the case but it is speculated that the terms at such a level belong to a basic level category, the most common level of detail in categorisation. Shapes of hypernym chains become more complicated as their length increases, with shorter chains forming a monotonically decreasing line (e.g. length 3 and 4), middle chains (e.g. 5,6,7) having a peak in the middle, and longer chains (e.g. 8 to 15) having more than one peak.

Second experiment: Parent-child pairs

The second experiment investigated the number of cases where parent synsets held a higher document frequency than their children. It was also anticipated that this experiment would reveal the naïve performance of document frequency to identify a parent term for a given pair. The results of the analysis show that document frequency was most reliable to identify more general terms from given pairs if both terms were very specific. The overall performance for 26,678 distinct pairs in WordNet was 76%, which is a significant result.

Third experiment: Effect of co-occurrence information

The third experiment examined the impact of the co-occurrence of terms in sets of documents when considering term specificity. The assumption here is that the more documents a given pair of terms share, the more the two terms are related. 100 queries from TREC-6 and TREC-7 were used to retrieve 500 top ranked documents from each

of the Financial Times, LA Times and the Wall Street Journal TREC collections. Each retrieved set was analysed to identify the occurrence of WordNet synsets, as well as those pairs which were found to collocate in the documents. For each recorded pair the numbers of cases where parent terms held a higher document frequency than the child were also recorded. Document frequency with and without co-occurrence information was thus measured in three corpora: the 500 top documents, a super-set of TREC newspaper collection and Web pages indexed by Google. The results show that the Web based larger size corpora was more accurate at determining specificity. It was also easier to determine specificity from co-occurring term pairs and the impact was more significant in the smaller corpora. Although the corpora were different, the usage of commonest sense as defined by WordNet was similar across all the corpora. However, the coverage of the WordNet vocabulary differed, with Google including 99%, whereas the TREC collections covered just over 50%.

The three experiments are reported in a draft paper, Joho, H, Sanderson, M and Beaulieu, M. Large scale testing of document frequency's relatedness to term specificity. To submitted to *Information Retrieval*. See Appendix 4.

4.2 Phase 2: User validation of concept labels and relationships

The second phase of the project was concerned with evaluating the adequacy and performance of the concept structures for searching and retrieval from a user perspective. The object was twofold: firstly to test how users interpreted concept labels in a query expansion task, and secondly to test to what extent concept structures provided an overview of the retrieved document set. The formative user test thus compared two methods of presenting candidate expansion terms, a hierarchical presentation and a conventional list. For the experimental system candidate expansion terms were first extracted in advance from the top 500 documents retrieved by the INQUIRY retrieval engine, in response to a query compiled from terms in the title of 45 TREC6 topic descriptions, then organised by the subsumption process, and finally visualised by the cascade menus. Lists were also generated using the identical set of terms included in the menus and ordered alphabetically. Twenty-four test subjects were asked to select terms they deemed appropriate to expand nine given queries. The experimental group was presented with an interface containing menus and the control group was presented with lists. Following the completion of the expansion tasks by all the participants, selected expansion terms were added to the initial query for each topic and the search was re-run and results compared.

Results show that the unexpanded queries retrieved documents at a higher precision for lower recall whilst the expanded queries produced higher precision for higher recall. Although queries expanded by the concept menus performed slightly better than the lists, the difference was not statistically significant. Nevertheless both approaches led to new relevant documents to be retrieved in the top 20. Additional measures were also considered, namely the number of expansion terms selected by subjects and time taken to complete the expansion task. The Menu group completed the task in less time and with over four terms fewer than the List group on average and both performance measures were statistically significant. Moreover 80% of subjects in the Menu group indicated that they had a better idea of the contents of the retrieved documents through the concept hierarchies. Thus the approach seemed to provide an adequate summarisation of documents retrieved in response to a query and

the subsumption hierarchies were meaningful. A further analysis to determine the semantic and/or topic relations between initial query terms and expansion terms also showed that more than half of the expansion terms selected by subjects were conceptually related and were broader than related terms normally found in a thesaurus. The evaluation study was reported in. Joho, H, Coverson, C, Sanderson, M. and Beaulieu, M. Hierarchical presentation of expansion terms. *In: Proceedings of the 17th ACM Symposium on Applied Computing, Madrid Spain, 2002*. New York: Association for Computing Machinery, 2002, 645-649. See Appendix 5.

4.3 Phase 3: Test system configuration

The third phase of the project concentrated on the development of the system. Two major tasks were undertaken: the design of the interface to integrate the concept visualisation tools, i.e. subsumption hierarchy and the keyphrase hierarchy, and the implementation of the back end retrieval system

4.3.1 Integration of the concept visualisation tools into a user interface

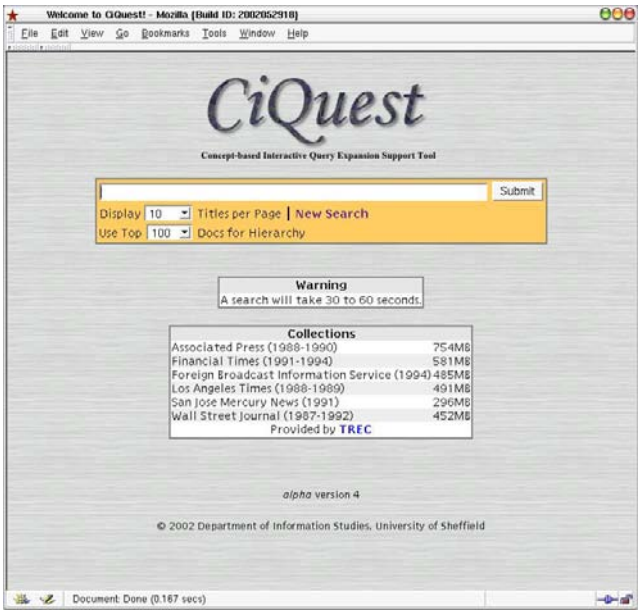
Although it was anticipated in the original proposal that a stand-alone front-end would be built using Java GUI components, it was decided that a web-based implementation of the system would be more appropriate. This is largely because most people access search engines through web-browsers. In addition the results of the formative evaluation of cascading menus also confirmed that this approach was effective. Hence Directed Acyclic Graph tools were not considered as a viable approach to pursue as had been anticipated at the outset of the project.

In implementing the concept hierarchies two main design issues had to be considered, namely the size of hierarchy to be displayed and how users will interact with the display. In theory a hierarchy can be very deep (or long) covering all the identified parent-child relationships. The width and depth of menus generated from a query can be seen as a characteristic of the set of retrieved documents. However the hierarchy containing too many terms could be confusing to users. Informal discussions with different users to consider the size of menus did not lead to any obvious solution and it was thus decided to limit the length to 20 layers in the actual test system and to 5 for demonstration purposes as illustrated below.

With regard to designing the interaction with the hierarchies, a major change was made following the formative evaluation conducted in phase 2. In the initial system, a term was simply added to a list when a user clicked a term from the menus. In the integrated system, a cascaded menu was shown along with the hitlist generated in response to a query. When a user clicks a term in the menu, a subset of documents in which the term occurs is shown.

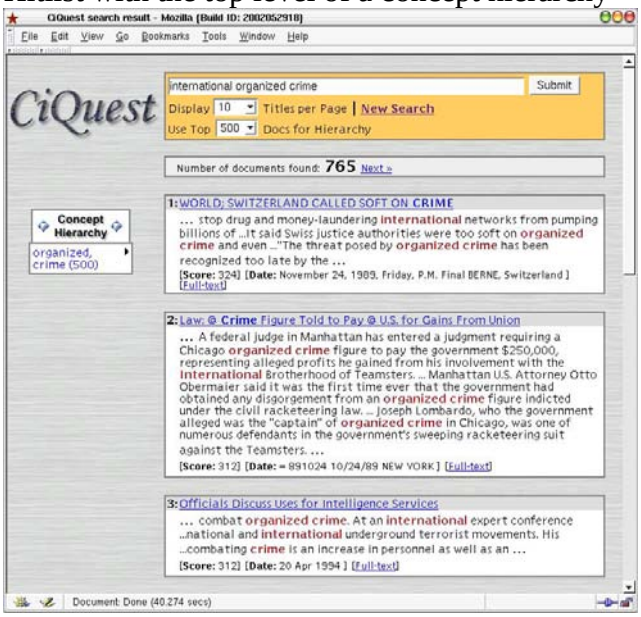
The following are screenshots of the demo system, which was demonstrated at the SIGIR 2002 conference in Tampere Finland. This offered another opportunity to obtain input from users and resulted in some very positive feedback. Further details are found in: Joho, H, Sanderson, M and Beaulieu, M. Hierarchical approach to term suggestion device. In*: Proceedings of the 25th Annual ACM SIGIR Conference on Research and Development in Information Retrieval, Tampere, Finland. 2002*. New York: Association for Computing Machinery, 2002, p.454. See Appendix 6.

Main Window

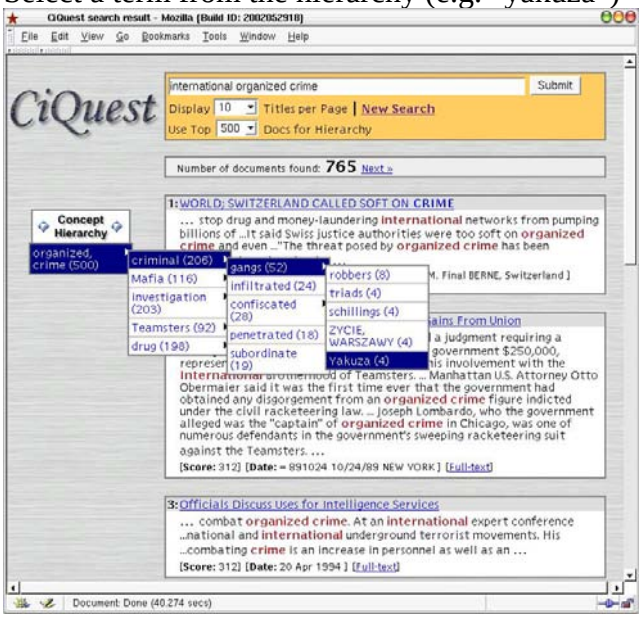


Query Box (Query: international organized crime)

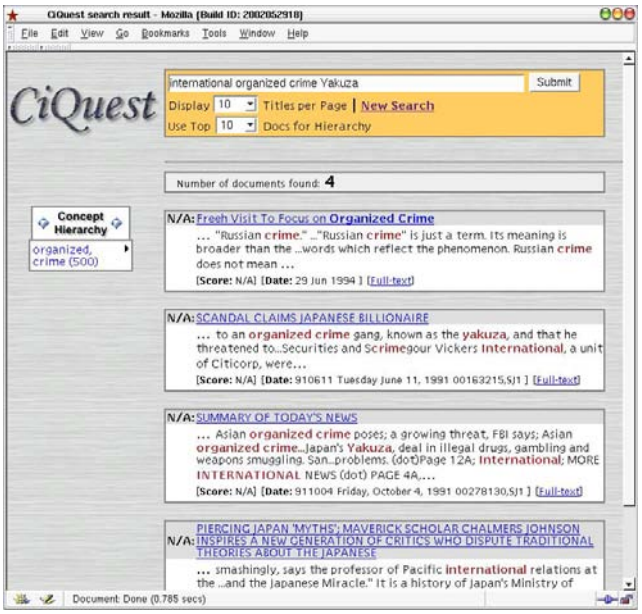
Hitlist with the top level of a concept hierarchy



Select a term from the hierarchy (e.g. “yakuza”)



A set of documents in which “yakuza” occurs was shown as a focused set of documents.



In the full test system two different approaches were tested to generate a concept hierarchy. Both approaches used the top 200 documents retrieved by a retrieval system in response to a query. Words and noun phrases were extracted from the documents using an internally developed part-of-speech tagger. The next two subsections will describe the two techniques for generating concept hierarchies. The example hierarchies are based on the query “Hubble Telescope Achievements”.

Subsumption hierarchy

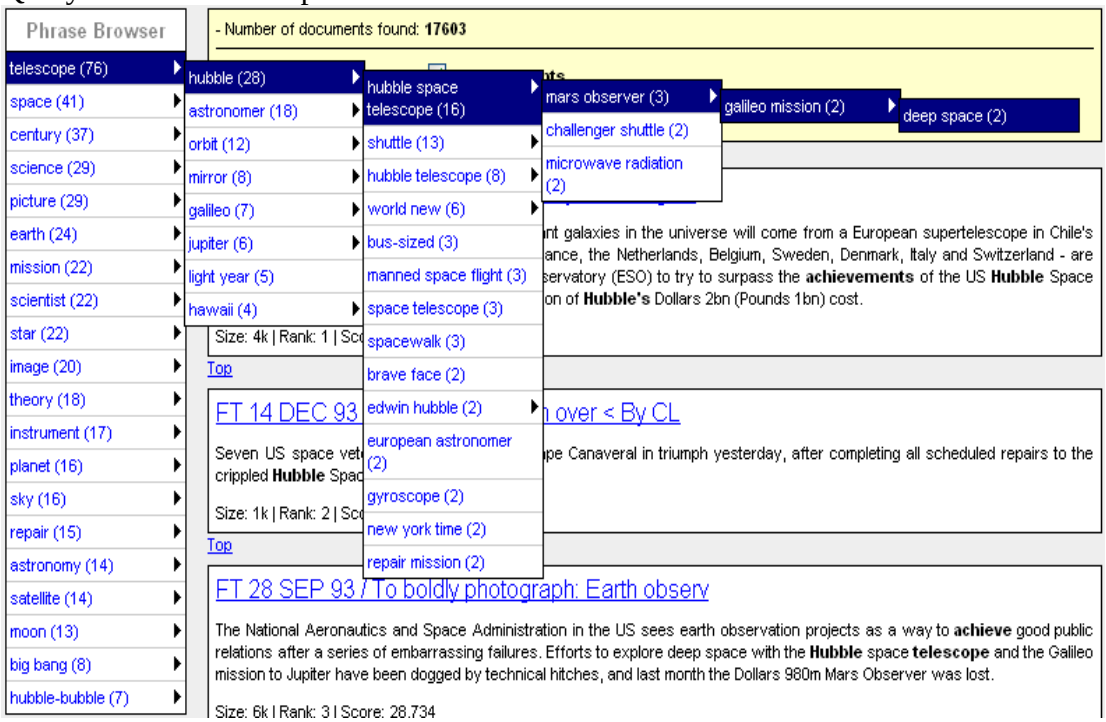
This type of hierarchy is based on statistical properties of terms, and called a subsumption hierarchy. The subsumption hierarchy determines which is more general

for a given pair of terms by analysing their document frequency and co-occurrence information. More specifically, a term X is said to subsume term Y when the following condition is held:

1) p(X|Y) >= 0.8; and
2) p(Y|X) < 1

In other words, when the majority of a set of documents in which term Y occurs is found to be a subset of documents in which term X occurs, term X is located in a parental position of term Y in a hierarchy. Since they co-occur in a reasonable amount of documents, they can be seen as related terms. Since term X occurs more frequently than term Y and term X's document sets subsumes term Y's, the former can be regarded more general than the latter.

Query: Hubble Telescope Achievements



Keyphrase hierarchy

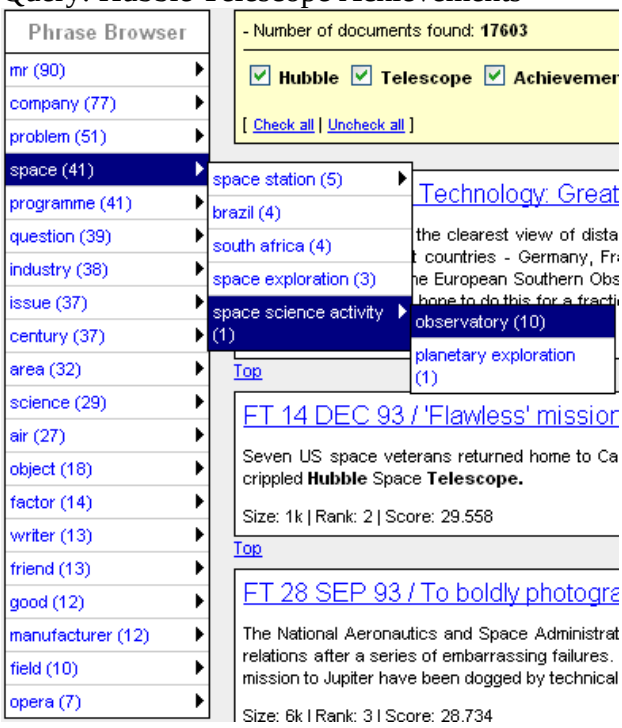
The second type of hierarchy is based on linguistic attributes of terms, and called a keyphrase hierarchy. The keyphrase hierarchy attempts to extract a parent-child pair from texts using a set of text patterns (referred as trigger phrase in this report). Some examples of the trigger phrases and the fragments of matched sentences are shown as follows.

1) NPa SUCH AS NPx, NPy, and NPz
"… international organisations such as NATO, WHO, and UNESCO are …"
2) NPx, NPy, NPz AND OTHER NPa
"… Google, Altavista, Yahoo and other search engines are …"
3) NPa, ESPECIALLY NPx, NPy, and NPz

where NP stands for a noun phrase.

As can be seen the trigger phrases are designed to find a sentence that have a noun phrase (i.e. NPa) and its instance or member (i.e. NPx, NPy, and NPz). When such a pattern is found, NPa is located in a parental position of NPx in a hierarchy. Also the decomposition of head nouns (e.g. organisations of “international organisations”) is applied so that a head noun becomes a parent of its noun phrases.

Query: Hubble Telescope Achievements



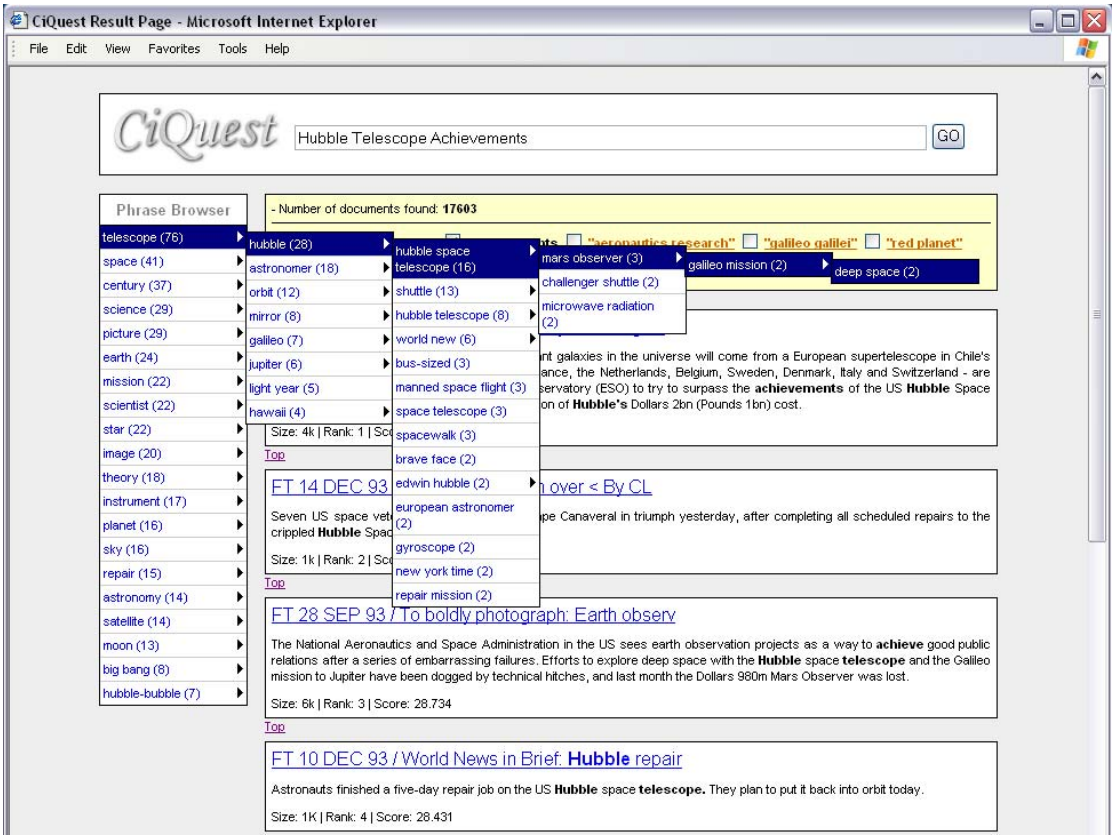
Selected Terms

In addition information about the terms selected from a menu was added to the full test system. When a box besides a term is ticked, the term is added to the query box. When a term itself is clicked, a list of documents linked to the term is shown.

Other minor changes

A couple of slight changes were made on the interface after the demo-system. For example, a box that contained some system parameters was removed from the final system. Also the general colour scheme was changed to a mono-tone style so that every keyword was well highlighted with a different colour set. In addition the name of the hierarchy was changed to “Phrase Browser” as we felt this would be more meaningful.

The screenshot of the final system is found below.



4.3.2 Back-end system

In the demo, an internally developed IR system was used as a back-end of the CiQuest system. However in the final implementation, the Okapi system was adapted since the previous system did not have a best-passage function. The best passage was displayed for each record in the hitlists.

A session manager component was also added to the system to record users' search activities. Information such as a time stamp, submitted queries, selected terms, displayed and/or saved documents, and paging information were logged by the session manager.

4.4. Phase 4: Summative user evaluation of the test system

The last phase of the project was to undertake a summative evaluation of the system to ascertain the retrieval effectiveness and usability of concept structures and the keyphrase finder in supporting interactive query expansion. The interactive tools incorporated into the CiQuest system adopted two different approaches. On the one hand the generation of concept hierarchies derived from term subsumption provided a form of summary and a way of visualising a retrieved document set. On the other hand, the Keyphrase finder extracted useful keywords from the retrieved documents that could be presented to the user as useful suggestions for modifying and improving their original queries. Although both tools were very different in the way each functioned, their purpose was the same, i.e. to enhance query formulation and interactive searching.

The experiment was based on the CiQuest system as described above but three different versions were devised for the test. The first was a baseline system which offered no query expansion tools. The second and third versions each incorporated the subsumption tool and the Keyphrase Finder respectively. Although the underlying functionality was different, subjects were not made aware of this as they searched through a common web-based interface. Twelve test subjects were recruited to undertake searches on three TREC-8 topics, one on each of the system test versions. Participants were briefed to undertake two tasks. The first was an `instance' finding tasks as developed by the TREC Interactive Track which entailed finding as many different answers to a topic or question as possible with the object of saving at least one document for each of the different aspects or answers of the topic. 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 rather than instance relevance. Subjects filled questionnaires at the beginning of the test session, after each search, and on completing the whole experiment. The procedure took 60 to 90 minutes for each subject.

The main results indicate that optimised queries performed better than initial queries, that is the number of relevant documents retrieved was statistically significant. Although overall changes to initial queries were small, these nevertheless lead to improved effectiveness. The keyphrase finder led to higher precision followed by the baseline system and the subsumption tool but this was not statistically significant. Whilst automatic query expansion normally works better when the initial query performs well, in the present study retrieval effectiveness improved regardless of the performance of the initial query.

With respect to searching efficiency and searching behaviour, the subsumption and keyphrase tools led to subjects viewing fewer documents than the baseline but saving more documents. The number of query iterations for the two menu based systems was also lower than the baseline. The subsumption tool was used more and led to more saved documents. It would appear that it allowed searchers to focus on a smaller set of documents. With regard to user perceptions, the subsumption tool was considered easier to use than the keyphrase finder and the size of menus for each could be handled equally well. Both tools were also considered to be equally useful in predicting the `aboutness' of documents linked to menu options, in focusing on useful new terms, and in helping in the assessment of relevance. However the subsumption tool was deemed to be better for gaining an overview of the contents of individual documents as well as a document set.

The study also provided insight into browsing behaviour. When a subject found a reasonable amount of relevant documents in the first page of results, they tended to go on to the next page. But the subsumption and keyphrase tools were accessed more frequently when the first page of results was less satisfactory. The concept and phrase browsers were typically used in two ways. One approach and the most popular was to focus on a subset of documents of interest. Another approach was to quickly and repeatedly iterate between menu terms and the title and best paragraph of the linked documents. Finally browsing menus tended to be a very quick operation. There was no evidence that semantic connections between a parent and child term significantly

influenced browsing. This may explain why most subjects did not realise the difference between the two types of menus. Another observation was that the occurrence of query terms at the top level of the menu seems an important factor in encouraging subjects to use the term browsers and subjects tended to go back to a parent term when one of its children was found to be useful and try another child term. The evaluation study is reported in: Joho, H., Sanderson, M., and Beaulieu, M. (2004) "A Study of User Interaction with a Concept-based Interactive Query Expansion Support Tool" In: *Proceedings of the 26th European Conference on Information Retrieval*, Sanderland, UK: Springer. See Appendix 7.

5. Summary and conclusions

The CiQuest project successfully met the stated aims and objectives set out in the original proposal as well as achieved some additional outcomes as summarised below:

- The main findings demonstrate that concept hierarchies can effectively be generated from sets of retrieved documents and displayed to searchers in a meaningful way.

- The approach provides the searcher with an overview of the contents of retrieved documents which in turn facilitates the viewing of documents and selection of the most relevant ones.

- Concept hierarchies derived from retrieved documents are a good source of terms for query expansion.

- Query expansion based on concept hierarchies can improve precision as opposed to query expansion based on relevance feedback which tends to improve recall

- Cascading menus make it easy for searchers to browse concept hierarchies quickly not only for selecting terms for query expansion but also for viewing clusters of documents within a retrieved set in a more focused way.

- The extraction of descriptive phrases as a source of terms for query expansion is also effective in improving retrieval performance. The comparable results obtained in using concept hierarchies as well as key descriptive phrases indicate that both approaches are equally useful.

- The project demonstrated the value of different approaches to test collections in conducting both small and large-scale retrieval tests based on TREC data as well as Web generated data.

- The project provided further insight into interactive searching behaviour and demonstrated the importance of user-based experiments in the development of interactive tools.

6. Dissemination and future work

The project has had an extensive dissemination programme with four contributions to major conferences being made in the course of the project. The next stage is to submit papers in refereed journals. Two papers have already been drafted and a third is being planned. The CiQuest system is also being made available to the IR research community through the Information Retrieval Research Group Website at Sheffield available at: http://ciquest.shef.ac.uk/~hideo/system/ (Access upon request).

CiQuest has had an impact on other work undertaken within the group. For example concept hierarchies have been incorporated in a cross-language retrieval system and experiments have been undertaken with cascaded menus for displaying terms in multiple languages.

A number of MSc projects have been carried in relation to CiQuest and it is envisaged that the system will form the basis for other studies. In particular there is scope to develop the system to incorporate different query expansion tools, i.e. concept hierarchies, descriptive phrases and relevance feedback. In addition larger scale user studies would be beneficial as well as more in-depth usability studies.

References

Evans, D A and Lafferty, R G. (1994). Design and evaluation of the CLARIT – TREC 2 system. In: D K Harman (ed). *The Second TREC Retrieval Conference (TREC-2),* NIST, 137-150.

Greffenstette, G. (1994). *Explorations in automatic thesaurus discovery*. Kluwer Academic Publishers.

Jensen, B J, Spink, A, Bateman, J and Saracevic, T. (1998). Searchers, the subjects they search and sufficiency: a study of a large sample of EXCITE searches. *Proceedings of WebNet'98, Orlando Florida, 1998.*

Jones, S, Gatford, M, Hancock-Beaulieu, M, Robertson, S E, Walker, S and Secker, J. (1995). Interactive thesaurus navigation: intelligence rules OK? *Journal of the American Society of Information Science*, 46(1) 52-59.

Kristensen, J. (1993). Expanding end-users'query statements for free text searching with a search-aid thesaurus. *Information Processing and Management*, 29(6), 733-744.

Robertson, S E, Walker, S and Beaulieu, M. (1997). Laboratory experiments with Okapi: participation in TREC. *Journal of Documentation*, 53(1), 20-34.

Sanderson, M. and Croft, B. (1999). Deriving Concept Hierarchies from Texts. *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 206-213, Berkeley, CA, ACM.

Appendices

- **Appendix 1:** Joho, H. and Sanderson. Retrieving descriptive phrases from large amounts of free text. In: *Agah, A. Callan, J & Rundenstenier, E. (eds). Proceeding of the 9th International Conference on Information and Knowledge Management,* McLean, Virginia, 2000. New York: Association for Computing Machinery, 2000, 180-186.

- **Appendix 2:** Joho, H., Liu, Y.K. and Sanderson, M. Large scale testing of a descriptive phrase finder. *In: Allen, J. (ed). Proceedings of the 1st Human Language Technology Conference,* San Diego, California, 2001. Morgan Kaufmann, 2001, 219-221.

- **Appendix 3:** Joho, H, Liu, T. T., Liu, Y. K., Sanderson, M, and Beaulieu, M. Descriptive phrase finder: a free-text mining tool for descriptive information about proper nouns. To be submitted to the *Journal of the American Society for Information Science and Technology.*

- **Appendix 4:** Joho, H, Sanderson, M and Beaulieu, M. Large scale testing of document frequency's relatedness to term specificity. To be submitted to *Information Retrieval*.

- **Appendix 5:** Joho, H, Coverson, C, Sanderson, M. and Beaulieu, M. Hierarchical presentation of expansion terms. *In: Proceedings of the 17th ACM Symposium on Applied Computing,* Madrid Spain, 2002. New York: Association for Computing Machinery, 2002, 645-649.

- **Appendix 6:** Joho, H, Sanderson, M and Beaulieu, M. Hierarchical approach to term suggestion device. In*: Proceedings of the 25th Annual ACM SIGIR Conference on Research and Development in Information Retrieval,* Tampere, Finland. 2002. New York: Association for Computing Machinery, 2002, p.454.

- **Appendix 7:** Joho, H, Sanderson, M and Beaulieu, M. A study of user interaction with a concept-based interactive query expansion support tool. In*: Proceedings of the 26th European Conference on Information Retrieval,* Sanderland, UK. 2004. Lecture Notes in Computer Science, Berlin: Springer-Verlag, 2004.

Appendix 1

Joho, H. and Sanderson. Retrieving descriptive phrases from large amounts of free text. In: *Agah, A. Callan, J & Rundenstenier, E. (eds). Proceeding of the 9th International Conference on Information and Knowledge Management, McLean, Virginia, 2000*. New York: Association for Computing Machinery, 2000, 180-186.

Retrieving Descriptive Phrases from Large Amounts of Free Text

Hideo Joho
University of Sheffield
Western Bank, Sheffield
S10 2TN, UK
+44 (0)114 222 2675
h.joho@sheffield.ac.uk

Mark Sanderson
University of Sheffield
Western Bank, Sheffield
S10 2TN, UK
+44 (0)114 222 2630
m.sanderson@sheffield.ac.uk

ABSTRACT
This paper presents a system that retrieves descriptive phrases of proper nouns from free text. Sentences holding the specified noun are ranked using a technique based on pattern matching, word counting, and sentence location. No domain specific knowledge is used. Experiments show the system able to rank highly those sentences that contain phrases describing or defining the query noun. In contrast to existing methods, this system does not use parsing techniques but still achieves high levels of accuracy. From the results of a large-scale experiment, it is speculated that the success of this simpler method is due to the high quantities of free text being searched. Parallels between this work and recent findings in the very large corpus track of TREC are drawn.

Keywords
Information retrieval, descriptive phrase, large corpora.

1. INTRODUCTION
The opportunities to use online text databases for the mining of valuable information are great. As these stores increase in size, the possibility of accurately extracting that information using increasingly simpler techniques seems to also increase. This principle was demonstrated in the results of the Very Large Collection (VLC) track of TREC-6 [4]. In the track the same topics as those used in the ad hoc task were applied to a 20Gb collection, which is a superset of the standard collection (2Gb). When comparing the effectiveness of systems retrieving on the two collections, it was noted that precision measured at rank position twenty was consistently higher for the systems searching the larger VLC. The reason for this was not explained by differences in retrieval techniques between the two runs, but that in the VLC, there were simply more relevant documents that held a high percentage of the specified query terms. In other words, because the collection was larger, users had a better chance of finding relevant documents that used the same combination of words and phrases as found in their query[1]. This effect only occurs with high precision measures: when considering all relevant documents in the VLC, the retrieval systems were not performing better, just a greater fraction of relevant documents were appearing in the top ranked positions.

The result from VLC implies that for retrieval tasks where finding a small fraction of relevant items is more important than finding them all, use of large corpora and simplistic matching techniques is likely to be a promising approach. Retrieving the descriptive or defining phrases of a proper noun is one such task.

Discovering the meaning of a particular word or phrase can be vital to the understanding a text. The conventional source of such a meaning (a dictionary) is often inadequate when the word in question is a proper noun. Other locations of reference information such as encyclopædia or online services, may be not easily accessible, have wide enough coverage, or be sufficiently up to date. Locating definitions within free text documents is an alternative approach[2].

A noun phrase defining or describing another noun within the same sentence is known as an *appositio*n. For example in the paragraph above, "encyclopædia" is described in the same sentence by the phrase "locations of reference information". While not a perfect or complete definition, it nevertheless provides some information on the meaning of the term. Finding this information within free text may at first seem to be a hard problem, however, work in a related area has shown that descriptions worded in a certain way can be located.

Hearst studied the problem of locating the IS-A lexical relationship within corpora [6][3] (taking an example from her paper, a broken bone *is an* injury). She showed that a word and its hypernym (the more general term) were often found together in sentences linked by common phrases; she manually and then semi-automatically located patterns that were reliable indicators of

[1] This will only work if the query does not match well to non-relevant documents.

[2] Motivation for this work came after one of the authors moved to the US and encountered a large number of references, in documents or in conversations, to proper nouns specific to US culture, which were not easily found in reference works.

[3] Hearst originally reported this work in an earlier publication [5].

the IS-A relation. These *key phrases* were "such as", "and other", "or other", "especially", and "including". Hearst looked for these phrases in the following patterns.

- *(dp* such | such *d*p) as *qn*
  - e.g. "...injuries such as broken bones"
- *qn* (and | or) other *dp*
  - e.g. " broken bones and other injuries..."
- *dp* especially *qn*
  - e.g. "... injuries especially broken bones"
- *dp* including *qn*

Here, *qn* (a noun later referred to as the *query nou*n) is the hyponym, and *dp* (a noun phrase later referred to as the *descriptive phras*e) is the hypernym.

Hearst reported on the accuracy of the phrases and discussed using the large number of "IS-A" relations listed in WordNet [8] to try to find other such indicative phrases. However, she was unable to find a fully automatic method for locating them.

Hearst stated that her technique "…is meant to be useful as an … aid to lexicographers…". Her work, however, has a broader applicability: the hypernym of a word appearing in the same sentence of that word is an apposition. Therefore, the key phrase method can be used as a starting point for building a system to locate descriptions of nouns. The coverage of such a method is likely to be poor, as the key phrases are relatively rare. Users of such a system are unlikely to want to see all descriptions (preferring high precision to high recall), as long as a few are found relatively accurately, most will be satisfied. Similar to the VLC track of TREC, as long as the corpus being searched is large enough, the likelihood of locating a description of the query noun in a "key phrase form" will hopefully be high.

It is with a strategy based in part on Hearst's lexical relation method that a descriptive phrase retrieval system was created. Starting with a review of previous work in the area, the rest of this paper describes the system's design, building, and testing. This is followed by speculation on possible future work and the paper closes with conclusions.

2. Previous work

The descriptive phrase retrieval system has similarities to some of the MUC (Message Understanding Conference) tasks and to the field of Question Answering (QA).

The annual MUC conferences of the 1990s tested research groups' abilities at various Information Extraction techniques [2]. The basic task was to fill a template with information extracted from a stream of documents. One of the slots in the template was used to hold any extracted description information of the entities identified in the texts. This task has clear similarities to the problem described in this paper. However, the methods used to process the documents in MUC were usually specialised to a particular domain making use of parsing technologies. Therefore, the solutions proposed were of less use as the aim of the work in this paper was to have a system as widely applicable as possible.

Interest in QA has long been active. Cooper described what he called a "fact retrieval system", which would search for text fragments in a small document collection which confirmed or denied a query statement [3]. Using a hand built parser working over a small set of sentences, Cooper reported some experimental success on his limited domain system. With a more general approach, Kupiec described a system that searched an online encyclopædia for answers to a set of closed class questions (in this work only "Who..." and "What..." questions were fully implemented, e.g. "Who's won the most Oscars for costume design?") [7]. The system used a parser to locate and type the important phrases within a question. From this information Boolean queries were constructed to search for sentences in the encyclopædia. These sentences were themselves parsed to try to find potential answer phrases. Secondary queries were then constructed to try to confirm which, if any, of these identified phrases was a valid answer. For example, "Who..." questions were expected to have a person's name as an answer, for these, the secondary queries were used to confirm a potential answer was actually a name. Kupiec tested his system on seventy questions taken from the board game Trivial Pursuit. The highest ranked sentence returned by the QA system for each query was correct 53% of the time. Within the top five sentences, the correct answer was found in 74% of the questions. More recently, there has been a growth of interest in QA with the Question Answering track of TREC. According to [13], two of the better performing systems in the TREC-8 evaluation were from [9] and [12], both of who made extensive use of parsers, existing knowledge bases and pre-calculated question templates. The differences between the work of this paper and QA are described in the next Section (2.1).

Work on explicitly extracting descriptive phrases was recently conducted by Radev [11]. His system was presented with a user specified query noun and it would locate and return a list of descriptive phrases of that noun extracted from a database of news web sites. Although it is not described in detail, it would appear that the system used an grammar to locate one of two basic syntactic patterns and their variants:

- *dp qn*
  - "Politician Tony Blair ..."
- *q*n, *d*p, or *d*p, *q*n
  - "Tony Blair, politician, ..."

The system was also capable of typing the descriptive phrases, deciding for example if the phrase was a location, an occupation, an age, etc. After manually examining 611 descriptions identified by the system, Radev found that they were correct 90% of the time. No results on the accuracy of the typing of descriptions were reported.

2.1 Design of the system

Question Answering is a more general problem than the locating of descriptive phrases. A system performing this more restricted task can be thought of as a specialised QA tool capable of answering the questions "Who is *q*n?" and "What is *q*n?". There are advantages to be had by concentrating on this smaller problem. First, as will be seen below, solutions to this particular sub-problem of the QA task perform well without use of specialised domain knowledge or language tools and so can be

expected to operate in a wide range of domains with little or no adjustment. The second advantage stems from the answers expected for these particular problems. In the wider QA task, one cannot assume how often the answer will occur in the collection to be searched. When searching for the descriptive phrases of a query noun, however, it is believed that there is a greater likelihood of the descriptions appearing many times across many documents and, as will be seen, this abundance of answers can be exploited in ways perhaps less used in QA.

The design of the system was as follows, given a query noun *(q*n), all documents holding it were retrieved and from them all sentences containing *qn* were extracted. These were ranked based on a series of criteria described below. We evaluated the top five and top ten highest ranked sentences for relevance. The system was judged successful for a particular query if at least one sentence in the ranked list contained information answering, at least in part, the who or what question. It may seem that this is a rather low measure of success, however, it is believed that in this task, users will be more than capable of locating the real description and ignoring the other non-relevant sentences.

Three criteria were used to rank the sentences:

- presence of a key phrase in the sentence,
- a high number of common terms,
- and the position of the sentence as found in the document.

Each of these features is now described, followed by the means of their combination.

2.2 Key phrases

Key phrases were used as the basis of the detection system. Using the phrases already listed in Section 1, three more were added: one to find acronyms, one to find "is a" type descriptions and another to locate appositions parenthesised by commas (similar to [11]). The patterns were defined as follows

- *qn (d*p) or *(d*p) *qn*
  - e.g. "MP (Member of Parliament)"
- *qn* (is | was | are | were) (a | an | the) *dp*
  - e.g. "Tony Blair is a politician..."
- *q*n, (a | an | the) *d*p
  - e.g. "Tony Blair, the politician, ..."
- *q*n, which (is | was | are | were) *d*p,
- *q*n, (a | an | the) *dp* (. | ? | !)
- *q*n, *d*p, (is | was | are | were)

The system to match these patterns required approximately seventy lines of Perl script. In contrast, both [6] and [11] used parsers to process candidate sentences. In this work, however, it was decided to avoid the use of these more complex tools so as to examine how successful a technique based on simple pattern matching would be. If it proved to be just as effective, this approach would in all likelihood be preferable to a parser based method because of its speed, simplicity, and potential applicability to a wide domain.

In the system for this paper, it was judged that a set of ranked sentences should always be returned to the user regardless of the success of the pattern matching. It was quite possible that descriptions of the query noun were present but had not been found through a mistake or lack of coverage in the patterns being matched. Therefore two additional more general criteria were included. It was anticipated that they would act as both a fallback when no patterns matched and as a way of ranking sentences found to match a pattern hopefully ranking better descriptions higher. The criteria were based on the information retrieval (IR) related techniques of location within a document and cross-document term weighting and are now described.

2.3 Sentence position

It seemed reasonable to expect that if a noun was used a number of times within a document, then any accompanying description of it was going to be found nearer the start than the end. Therefore the ordinal position of sentences containing the query noun (e.g. 1st, 2nd, 3rd, etc) was noted and used in the ranking calculation. The earlier sentences were given a higher score.

2.4 Word counting method

If a query noun was described in one document, it was assumed that it was likely to be described in others. It was anticipated that this repetition of the same or similar descriptions across documents could be exploited. A simple word counting technique was devised to examine all sentences retrieved in response to a query noun and to find words co-occurring with the noun that commonly co-occurred across documents. A number of methods were tried, the one found to be most successful (evaluated on a test collection described below) involved examining in each document (containing *q*n) only the first sentence that *qn* occurred in. The case of the words in these sentences was normalised, stop words were removed, and a stemmer was applied [10]. The frequency of occurrence of all remaining terms in the sentences was calculated and the twenty most frequent were noted. When ranking all matching sentences, each was assigned a score based on the number of the top twenty terms present. Those containing more of these common terms were given a higher score.

2.5 Tuning and combining the criteria

Before any evaluation of the system could take place, it was necessary to tune it to try to get an optimal performance from the three sentence ranking criteria. Therefore, a descriptive phrase test collection was created half of which (the training set) was used for tuning the system, the other half (the test set) used later for evaluation. This section describing the building of the collection and its use in system tuning.

2.5.1 Building the collection

The document collection to be searched was a set of LA Times articles from 1989 & 1990 (475Mb, taken from TREC). The advantage of using the TREC data was the relative ease of access to it. The disadvantage was that the authors were left with the challenge of thinking of a large number of query nouns that might have been described ten years ago. Seventy six such queries were thought by the authors (or suggested by colleagues) of which ten were not present in the collection and a further sixteen that only

occurred a small number of times (<20). These final sixteen were removed, as it was felt that there was little challenge in finding those sentences that described them as a user would probably be willing to read through all the sentences retrieved by those queries. The remaining fifty queries were used in the test collection. They occurred in 16,111 sentences; each of these was assessed for relevance. As stated above, a sentence was judged relevant to a query if it contained information that would help a user understand more about the noun they queried on. As with all relevance judgements there were some sentences that were hard to decide on. For example in one sentence containing the query "Adolf Hitler", he is described as a person.

"...not only for Kraft but for people such as Adolf Hitler and Adolf Eichmann..."

Although this is a valid description the sentence was judged to be not relevant. It is hard to imagine someone not knowing that a person's name refers to a person. This type of problem was, however, the exception and for most sentences it was clear if it contained a description or not.

Note that judging relevance was on sentences and not on extracted descriptions or some other unit of text smaller than a sentence. This made it possible to automatically process the results of the descriptive phrase system in a similar manner to traditional IR evaluation. This contrasts with the evaluation method used in the QA track of TREC, where judges had to examine each individual answer from each run of each participating system [13].

Evaluation was measured using TREC-like measures concentrating on the rank-based ones: precision at ranks 1, 5, and 10. In addition the number of queries for which at least one correct answer was found within those rank positions was also calculated.

2.5.2 Tuning key phrases

While writing the key phrase pattern matcher, it was clear that some of the patterns were going to be better at locating descriptive phrases than others. Therefore, the training set was examined to measure the accuracy of the patterns. Table 1 shows this along with their coverage.

As can be seen, all the patterns are relatively rare (compare with numbers for no patterns), though the comma parenthesised apposition and "such as" were the most used, "and other" proved to be most accurate. These figures were used in the ranking of sentences with those containing the more accurate key phrases getting a higher score.

2.5.3 Combining the criteria

A series of tests were run on different combinations of the scores gained from the key phrases, the sentence location, and co-occurring word counts. A weighted sum of the scores was found to work best

$$aKPW + bWC + c(d - SN)$$

where *KPW* is the key phrase accuracy weight taken from above, *WC* is the co-occurring word count, *SN* is the sentence number (1st, 2nd, etc sentences occurring earlier in documents got a higher score). The values of a, b, c, *d* (tuning constants) were set to 2000, 1, 75, and 500 respectively after a series of trials on the training set. In the trials, different combinations of the constants were examined, each time measuring the effectiveness of the system using the performance measures outlined in Section 4.

Table 1. Accuracy of key phrase pattern matcher

	Not rel	Rel	Total	Accuracy
No pattern	6424	872	7296	12.0%
especially	0	0	0	0.0%
qn, dp,	89	63	152	41.4%
is a	23	18	41	43.9%
including	20	17	37	45.9%
or other	1	1	2	50.0%
such as	59	59	118	50.0%
acronym	14	23	37	62.2%
and other	9	23	32	71.9%

3. The system in action

Working with the training set seemed to show that the system was producing reasonable results. To illustrate the following are the manually extracted descriptions taken from a few high ranking relevant sentences selected at random from the query set.

- John Lennon - '60s rock artist
- Tofu - [no phrase found in top 10]
- Hitachi - top manufacturer
- NEC - established portable computer company
- Nintendo - a 99-year-old Japanese firm
- Star Wars - defence program

As with the Hitachi query, sometimes the description is too general. But for the most part, these descriptions seem to be reasonable. If someone knew nothing about these query nouns, the descriptions here would give that person more information than they had before.

The type of description found within a corpus clearly depends on the audience that the corpus was written for and how much it is thought that they already know. The documents used in this work were ten-year old US newspaper articles. Unless the queries being searched have an American or international significance, it is unlikely they will be found in the corpus.

The shortest description found was one word describing Bob Dylan as an "artist", one of the longest was of US TV news presenter Diane Sawyer described as "the Grace Kelly of television - the perfectly groomed Ice Queen whose every gesture seems scripted".

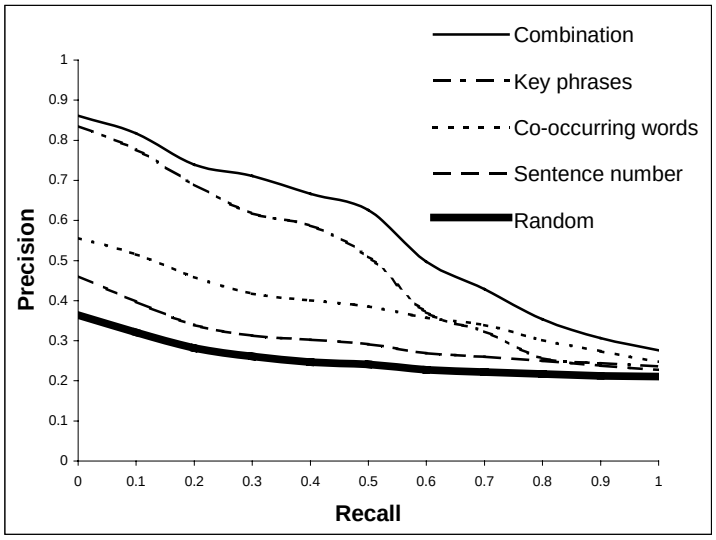


Figure 1. P-R graph of four strategies

4. Evaluation

Now that the system had been tuned, it was evaluated on the test set. Two experiments were conducted, the first was a test of the effect of collection size on the key phrase matching system, the second was a full evaluation of the system.

4.1 Collection size

In this experiment random samples of the test set were taken and the effectiveness of the key phrase criteria was measured for each of these samples. Results from this experiment are shown in Table 2. Samples taken were for 10%, 25%, 50%, 75% and 100% of the test set.

The results of this test show that as the collection got smaller, the effectiveness of the system reduced. This is of course because the likelihood of the system finding a sentence holding one of the key phrases reduced as the collection got smaller, therefore, the precision of the system in the top ranked sentences fell. This result echoes that obtained on the VLC collection described in Section 1.

Table 2. Precision (Key phrase) and collection size

P. at rank	100%	75%	50%	25%	10%
1	0.75	0.78	0.69	0.63	0.62
5	0.51	0.51	0.46	0.38	0.32
10	0.42	0.40	0.35	0.28	0.24

4.2 Testing the system

The effectiveness of each of the individual sentence ranking criteria was tested along with combination formula derived above. Unlike most IR test collections, the ratio of relevant to non-relevant was relatively high. Therefore, it was important to establish a random baseline as well. To achieve this, for each of the fifty queries, 100 random orderings of the sentence collection (in the test set) were generated and the average effectiveness of these cumulative runs was measured.

A precision-recall graph was plotted showing the effectiveness of the four strategies plus random retrieval (see Figure 1). As can be seen, the three criteria and their combination do better than random retrieval[4]. A sentence ranking based purely on the key phrase weights was extremely effective, except for high recall situations where the co-occurring word counting method was better. The most effective technique for finding descriptive phrases, however, was the combination formula, which, through a t-test, was found to be significantly better than any of the other methods, including key phrases. The difference between the combination and key phrase methods was found to be significant at $p<0.05$ for recall levels 0.1 and 0.2 and significant at $p<0.01$ for all higher values of recall. When evaluated with precision oriented measures a similar picture emerged. Table 3 shows precision

Table 3. Precision of each strategy

P. at rank	Comb.	Key phr.	Co-occ. Words	Sent. No.	Rand.
1	0.76	0.75	0.37	0.25	0.20
5	0.57	0.51	0.35	0.27	0.20
10	0.46	0.42	0.35	0.27	0.20
15	0.42	0.36	0.33	0.24	0.20
20	0.38	0.32	0.32	0.23	0.20
30	0.32	0.26	0.28	0.22	0.19
100	0.17	0.15	0.16	0.15	0.14

[4] The monotonically decreasing line of the random system is an artefact of the standard interpolation used when measuring precision at fixed recall levels. Although the relevant documents along the ranking are evenly (randomly) distributed, when precision is measured on this distribution the line shown in the figure is the result.

measured at rank positions ranging from one to one hundred.

As can be seen, the combination method is consistently higher than the key phrase. Like the precision recall graph, significance testing was performed: the combination method was found to be significantly better than key phrase for all rank positions from five through to 100 ($p<0.01$). As stated at the end of Section 2.5.1, the percentage of queries with at least one correct answer in the top *n* was also calculated. Here *n* was chosen to be 5 and 10 as it was felt that a user would be willing to look through this number of sentences. For the best performing method (combination) 90% of the queries had a correct answer in the top 5 (compared with 22% for random) and 94% correct in the top 10 (c.f. random 31%).

5. Conclusion

This paper has presented a means of locating descriptive phrases of a user specified query noun. A method designed to locate lexical relations within text, using key phrases, was applied to this new task. It was adapted by expanding the number of key phrases and by incorporating additional within document and cross-document information. More complex linguistic processing and reliance on lexical resources was avoided.

Through large-scale experimental testing, results showed the system was successful. In tests on a collection of over 8,000 sentences (the test set), the system was capable of ranking a description-bearing sentence within the top ten for 94% of the tested queries: a level of accuracy anticipated to be acceptable to most users.

The experiment confirmed earlier results from the TREC VLC track that showed that simple methods searching on a large corpus can produce accurate results.

6. Future Work

There are a number of possible areas of further work.

6.1 Extracting descriptions

The descriptions shown in Section 3 were manually extracted from the sentences. Currently the system can only present whole sentences to the user. Although the query noun and some of the words of the descriptive phrase can be highlighted, it would be preferable for the phrase to be automatically extracted. For sentences containing key phrases, a prototype extraction system has already been written. Although it has not been formally tested, it does appear to work well. No automatic method for extracting phrases from sentences where no key phrase was found has been created and this is something we plan to pursue.

6.2 Managing ambiguity and time

In Section 3 the descriptive phrase for the query "Star Wars" illustrates the ubiquitous problem of ambiguity (the term refers equally well to the defence program as it does to the science fiction film). Methods to classify a word into its different senses have been well researched and we plan to apply some of these techniques for this problem.

Related to ambiguity is the issue of time, the descriptions of things will alter: people will change their jobs for example. A means of detecting and presenting this change to users will also be explored.

6.3 Generality of descriptive phrases

From the examples shown in Section 3, it is clear that there are different levels of descriptions ranging from the general to the specific. We believe that it will be possible to estimate the generality or specificity of a description through use of a range of simplistic methods: using the description's inverse document frequency (or the idf of its component words) may provide an estimate of the specificity of the phrase. Use of this simple statistic has been used for this purpose before [1]. It may also be possible to examine the range of proper nouns that a particular description has been applied to and use this as a means of estimating generality of the phrase.

6.4 The web

Given that the system is designed to work best on very large corpora, the obvious VLC on which to apply the phrase description system to is the Web. We plan to use our system as a front end for an existing search engine (e.g. AltaVista) using the engine to retrieve relevant documents and then using our system to locate the descriptive phrases. We anticipate that this will further improve the accuracy of our simple yet effective system.

7. References

[1] Caraballo, S.A., Charniak, E., "Determining the specificity of nouns from text", *Proc. the joint SIGDAT conference on empirical methods in natural language processing (EMNLP) and very large corpora (VLC)*, 63-70, (1999).

[2] Chinchor, N.A., "Overview of MUC-7/MET-2", *Proc. the Message Understanding Conference Proceedings MUC-7*, (1998).

[3] Cooper, W.S., "Fact Retrieval and Deductive Question-Answering Information Retrieval Systems", *Journal of the AC*M, ACM Press, 11(2), 117-137, (1964).

[4] Hawking, D., Thistlewaite, P., "Overview of TREC-6 Very Large Collection Track", *NIST Special Publication 500-240: The Sixth Text REtrieval Conference (TREC 6*), E.M. Voorhees, D.K. Harman (eds.), 93-106, (1997).

[5] Hearst, M., "Automatic Acquisition of Hyponyms from Large Text Corpora", *Proc. the 14th International Conference on Computational Linguistics (COLING 92*), 539-545, (1992).

[6] Hearst, M.A., "Automated Discovery of WordNet Relations", *WordNet: an electronic lexical databas*e, C. Fellbaum (ed.), MIT Press, (1998).

[7] Kupiec, J., "MURAX: A Robust Linguistic Approach For Question Answering Using An On-Line Encyclopedia", *Proc. the 16th annual international ACM SIGIR conference on Research and Development in Information Retrieva*l, 181-190, (1993).

[8] Miller, G.A., "WordNet: A lexical database for English", *Communications of the AC*M, 38(11), 39- 41, (1995).

[9] Moldovan, D., Harabagiu, S., Pasca, M., Mihalcea, R., Goodrum, R., Girju, R., Rus, V., "Lasso: A Tool for Surfing the Answer Net", *NIST Special Publication XXX-XXX: The 8th Text REtrieval Conference (TREC 8)*, (1999).

[10] Porter, M.F., "An algorithm for suffix stripping", *Program - automated library and information system*s, 14(3), 130-137, (1980).

[11] Radev, D.R., McKeown, K.R., "Building a Generation Knowledge Source using Internet-Accessible Newswire", *Proc. the 5th Conference on Applied Natural Language Processing (ANLP*), 221-228, (1997).

[12] Singhal, A., Abney, S., Bacchiani, M., Collins, M., Hindle, D., Pereira, F., "AT&T at TREC-8", *NIST Special Publication XXX-XXX: The 8th Text REtrieval Conference (TREC 8)*, (1999).

[13] Voorhees, E.M., Tice, D.M., "The TREC-8 Question Answering Track Evaluation", *NIST Special Publication XXX-XXX: The 8th Text REtrieval Conference (TREC 8)*, (1999).

Appendix 2

Joho, H., Liu, Y.K. and Sanderson, M. Large scale testing of a descriptive phrase finder. *In: Allen, J. (ed). Proceedings of the 1st Human Language Technology Conference, San Diego, California, 2001*. Morgan Kaufmann, 2001, 219-221.

Large scale testing of a descriptive phrase finder

Hideo Joho
Department of Information Studies
University of Sheffield, Western Bank
Sheffield, S10 2TN, UK
+44 (0)114 222 2675
h.joho@sheffield.ac.uk

Ying Ki Liu
Department of Information Studies
University of Sheffield, Western Bank
Sheffield, S10 2TN, UK

Mark Sanderson
Department of Information Studies
University of Sheffield, Western Bank
Sheffield, S10 2TN, UK
+44 (0)114 222 2648
m.sanderson@sheffield.ac.uk

ABSTRACT
This paper describes an evaluation of an existing technique that locates sentences containing descriptions of a query word or phrase. The experiments expand on previous tests by exploring the effectiveness of the system when searching from a much larger document collection. The results showed the system working significantly better than when searching over smaller collections. The improvement was such, that a more stringent definition of what constituted a correct description was devised to better measure effectiveness. The results also pointed to potentially new forms of evidence that might be used in improving the location process.

Keywords
Information retrieval, descriptive phrases, WWW.

1. INTRODUCTION
Retrieving descriptions of the words and phrases, which are not often found in dictionaries, has potential benefits for a number of fields. The Descriptive Phrase Finder (DPF) is a system that retrieves descriptions of a query term from free text. The system only uses simple pattern matching to detect a description, and ranks the sentences that hold the descriptive phrases based on within document and cross document term occurrence information. The system does not attempt to extract descriptions from text, it simply locates sentences that are hopefully relevant to a user. It is assumed that users are able to read a sentence and locate any description within it. The advantage of using such an approach is that the DPF is much simplified and does not require parsing to find the exact location of the phrase. Due to its simplicity, it achieves a level of domain independence.

The DPF was implemented and succeeded in The DPF was implemented and succeeded in retrieving sentences holding descriptive phrases (DPs) of a wide range of proper nouns. Initial testing on a collection of LA Times articles from the TREC Collection showed that 90% of the queries had at least one correct DP in the top 5 ranked sentences and 94% in the top 10 ([3]). It was shown that the effectiveness of the system was in part due to the large amount of free text being searched. What was not shown by the experiment was if performance could be further improved by searching an even larger text. Consequently, a larger scale experiment was conducted, searching for phrases from the World Wide Web (WWW) using the output of a commercial Web search engine to locate candidate documents that were then processed locally by the DPF.

In addition to increasing the number of documents searched, more queries were tested and different definitions of relevance were tried. The rest of this short paper explains the system and shows the results of the expanded experiment, followed by pointers to future work.

2. THE SYSTEM
The Web-based DPF was composed of two parts: a front-end to an existing Web search engine, which fetched documents; and the system that located sentences holding descriptive phrases.

The Web front end simply routed queries to a Web search engine (Google), and the text of the top 600 documents returned by the engine was fetched, split into sentences (using a locally developed sentence splitter), and those sentences holding the query term were passed onto the DPF.

It ranked sentences on a score calculated from multiple sources of evidence. A detailed description of the DPF is found in [3]. The primary clue to there being a descriptive phrase in a sentence was the presence of a *key phrase* within it. An example key phrase was "such as", which may be found in the sentence: "He used several search engines *such as* AltaVista, HotBot and WebTop to compare the performance". If such a sentence were returned to a user who entered the query "WebTop", they would determine it was a search engine. Specifically, the DPF is searching for the key phrase in proximity to a query noun *(q*n) to locate a descriptive phrase *(d*p) e.g.

- ... *dp* such as *qn* ...

other key phrases used, some suggested by [2], were

- ... such *dp* as *qn* ...
- ... *qn* (and | or) other *dp* ...
- ... *dp* (especially | including) *qn* ...
- ... *qn* (*dp*) ...

- ... *qn* is a *dp* ...
- .. *qn*, (a | the) *dp*, ...

The phrases form the key part of the DPF as they identify well sentences likely to contain descriptions of *q*n. While the number of times a particular *qn* appears in a sentence with a key phrase are small, by searching a large corpus, like the Web, the chances of finding a few (accurately identified) descriptions of *qn* in the form required are high.

Based on results from a testing phase, certain key phrases were found more accurate at locating a descriptive phrase than others. Consequently, when ranking matching sentences, different scores were assigned depending on the accuracy of the key phrase found within. Since unfamiliar words tend to be explained or rephrased at the early part of a document, sentence position was also a factor in the rank score, with earlier sentences given preference. Finally, cross-document information was taken into account. Across all the matching sentences for a particular query, the occurrence of all the terms within the sentences was noted. It was anticipated that terms occurring more frequently within the set of sentences were likely to belong to descriptions.

Consequently, sentences holding a high number of commonly occurring words were given further preference in the ranking. The last two pieces of information not only improved the accuracy of ranking, but also enabled the system to produce reasonable results when no key phrases were matched. A training phase where the optimum balance between the sources of information was run on existing training data created from the LA Time corpus described in [3].

It may be reasonable to question why such a simple approach to extracting information from free-text sources be taken when more principled NLP-based techniques are well-established (e.g. [4], [5]). There are a number of reasons:

- Any simple approach is likely to be much faster than one that requires operations such as parsing.
- We believe that the use of simple but accurate methods searching over very large corpora provides a new means of determining lexical relations from corpora that are worthy of further exploration.

3. INITIAL STUDY

A pilot study was conducted, searching ten queries using the top hundred documents returned by Google. Of the ten queries, six had the best description located in the top two ranked sentences, two more queries had a good description in the top two. For all queries, a sentence holding a descriptive phrase was returned in the top five ranked sentences.

4. DEFINING RELEVANCE

In this and the previous evaluation described in [3], relevance was defined as a sentence that told the user anything about the query term: a liberal view of relevance (described here as *binary relevanc*e). The results from the pilot, under this interpretation, showed the system performed well. Consequently a more stringent form of relevance was devised. A sample answer for each query was solicited from users: for example, "the Prime Minister of Great Britain" for Tony Blair. Those *key answers* were taken as an acceptable criterion of highly relevant descriptive phrases. Sentences ranked by the system were then compared to the key answer. Correctness of DPs is not enough for this aim. Only a DP that described a query as well as a key answer was regarded as relevant. To illustrate, the sentence "Tony Blair is the current Prime Minister of the United Kingdom." was regarded as relevant, but "Tony Blair is a political leader" was not.

5. THE MAIN EXPERIMENT

A total of 146 queries were tested in the main experiment: 50 of which were evaluated based on key answers; 96 using binary evaluation. In the binary test, the DPF returned a relevant (descriptive) sentence in the top twenty sentences for all 96 queries. On average sixteen of the sentences returned were relevant to each query. The minimum number of relevant was six and maximum was twenty. Across the 96 queries, at least one relevant sentence was found in the top five for every tested query. This is a significant improvement over the previously reported experimental results where 90% of queries were answered in the top five.

Using more stringent key answer based relevance, the system succeeded in retrieving at least one relevant sentence in the top five for 66% of the queries, at least one in the top ten for 82%, and one in the top twenty for 88%.

These results show that the DPF searching the Web (1 billion documents) works dramatically better than the previous experiment using LA Times (100,000 documents). As was shown in previous work, the size of the collection impacts on the effectiveness of the system. This is because by searching a larger collection, there is a better chance of locating a relevant descriptive phrase in the format of one of the searched for key phrases. However in the previous work, there appeared to be an upper bound on the accuracy of the descriptive phrases alone. By searching a much larger collection it is speculated that the cross document term occurrence statistics used contributed significantly to improving the effectiveness of the system.

6. CONCLUSION

An existing descriptive phrase system was adapted to work with a Web search engine to locate phrases describing query words. The system was found to be highly effective at locating good descriptions: finding at least one high quality descriptive phrase in the top 10 returned sentences for 82% of test queries.

7. FUTURE WORK

We plan to undertake a number of further experiments, examining through tests, the ability of people to locate descriptions within the retrieved sentences. In addition, it was notable that the results of the full experiment were not as good as those from the pilot study. One difference between the two tests was the number of web documents examined: 100 top-ranked documents in the pilot; 600 for the expanded experiment. Given that a search engine generally retrieves more relevant documents in the higher ranks, there is likely to be more noise lower down. It is also significant that the search engine used was Google, which uses the *page rank* authority measure ([1]) to enhance its ranking. Therefore, we speculate that use of an authority measure

can be used to further improve the quality of our DPF. This will be investigated in future work.

8. REFERENCES

[1] Brin, S., Page, L. The Anatomy of a Large-Scale Hypertextual Web Search Engine, in Proceedings of the 7th International WWW Conference, April 1998, Brisbane, Australia.

[2] Hearst, M.A. Automated Discovery of WordNet Relations, in WordNet: an electronic lexical database, C. Fellbaum (ed.), MIT Press, 131-151, 1998.

[3] Joho, H., Sanderson, M. Retrieving Descriptive Phrases from Large Amounts of Free Text, in Proceedings of the 9th ACM CIKM Conference, November 2000, McLean, VA, 180-186.

[4] Radev, D.R., McKeown, K.R. Building a Generation Knowledge Source using Internet-Accessible Newswire, in Proceedings of the 5th ANLP Conference, March 1997, Washington, D.C., 221-228.

[5] Srihari, R & Li, W. A Question Answering System Supported by Information Extraction, in Proceedings of the 8th ANLP Conference, April-May 2000, Seattle, Washington.

Appendix 3

Joho, H, Liu, T. T., Liu, Y. K., Sanderson, M, and Beaulieu, M. Descriptive phrase finder: a free-text mining tool for descriptive information about proper nouns. To be submitted to the *Journal of the American Society for Information Science and Technology.*

Descriptive Phrase Finder: A Free-Text Mining Tool for Descriptive Information about Proper Nouns

Hideo Joho, Tsung-Te Lin, Ying Ki Liu,
Mark Sanderson*, and Micheline Beaulieu
Department of Information Studies, University of Sheffield,
Western Bank, Sheffield, S10 2TN, UK.
Tel: +44 (0)114 222 2648 **Fax:** +44 (0)114 278 0300
Email: m.sanderson@sheffield.ac.uk

Abstract

This paper presents the descriptive phrase finder (DPF), a system that retrieves descriptive phrases of proper nouns from free-text. Unlike existing question answering systems, the DPF was implemented using simple pattern matching and term occurrence information within and across documents. Three large scale experiments, using web search engines interfaced to the DPF, showed that such a straightforward approach can achieve high accuracy and coverage in locating the descriptive phrases of proper nouns. It was speculated that the success was due to the large amount of free-text being searched: firstly, because there was a better chance of locating text patterns indicative of descriptive phrases; and second, more text results in better term co-occurrence statistics. Differences between the three search engines were also examined, focussing on collection coverage and the ranking algorithms used. Both aspects were found to impact on system effectiveness.

*Corresponding author

1 Introduction

This paper presents a system, the Descriptive Phrase Finder (DPF), that retrieves descriptions of a query term from free-text, and reports three large scale experiments conducted using the system. A descriptive phrase is a phrase that explains or describes a word/noun phrase (referred as query). An example of descriptive phrases is "the world's largest PC software publishing house" for a query term *Microsoft*. The system only uses simple pattern matching to detect a description, and ranks the sentences that hold the descriptive phrases based on within document and cross document term occurrence information. The system does not attempt to extract descriptions from text, it simply locates sentences that are hopefully relevant to a user. It is assumed that users are able to read a sentence and locate any relevant description within it.

It may be reasonable to question why such a simple approach to extracting information from free-text sources be taken when more principled question-answering (QA) related techniques using natural language processing (NLP) are well-established (e.g. Kupiec, 1993; Radev and McKeown, 1997; Srihari and Li, 1999). There are a number of reasons:

- Any simple approach is likely to be much faster than one that requires operations such as parsing.
- We believe that the use of simple but accurate methods searching over very large corpora provides a new means of determining lexical relations from corpora that are worthy of further exploration.
- Due to its simplicity, it is anticipated that it achieves a level of domain independence.

The present study was also motivated by the findings from the Very Large Collection (VLC) track of TREC (Hawking and Thistlewaite, 1997) where it was found that IR systems produced higher values of precision at rank 20 when searching on the VLC as opposed to the smaller TREC ad hoc collection. The reason for this effect was simply due to there being a better chance of locating in the VLC (as opposed to smaller collections) relevant documents that matched the queries well. When searching for descriptive phrases simple methods exist that while being highly accurate have poor coverage. By searching a sufficiently large collection, such coverage problems will hopefully be mitigated.

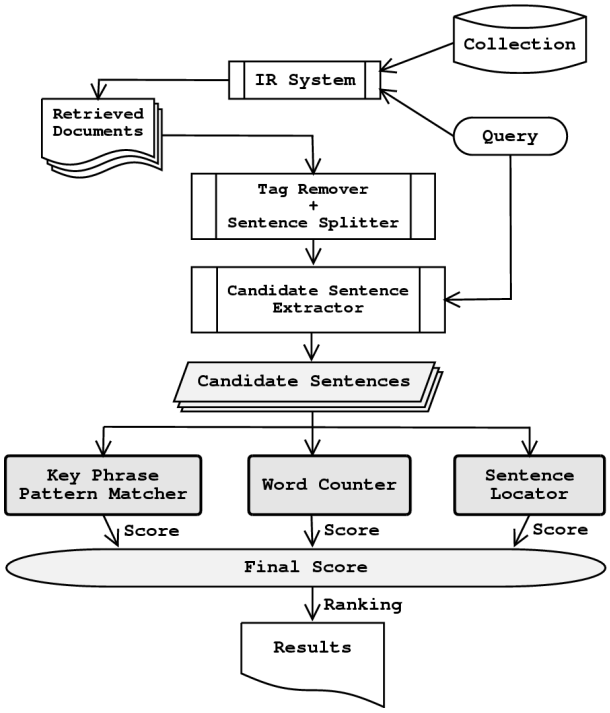


Figure 1: Overview of the DPF system

Previous studies (Joho and Sanderson, 2000; Joho *et al.*, 2001) have shown that the DPF system accurately locates descriptive phrases of proper nouns. This paper will include the major findings from the two previous works and then present a new experiment and results exploring the effects of using different Web search engines with the DPF. The next section will present the DPF system and its components, followed by the descriptions of the three experiments conducted on the system. A comparison will be then made with a system that is similar to the DPF system but with a more complex structure. Finally, conclusions and future work will be discussed.

2 The DPF System

This section presents the components of the DPF shown in Figure 1. The design follows the approach used in most large-scale QA systems where an IR system searches a collection in response to a query, passing a sub-set of documents onto an answer identification system.

2.1 IR System

An IR system internally developed by one of the authors was used for the first experiment, and a range of Web search engines were used for the second and third.

2.2 Pre-process components

The pre-process components consist of a tag remover, sentence splitter, and candidate sentence extractor. The tag remover strips markup tags from the documents retrieved by the IR system. The sentence splitter divides the texts into sentences. Of those sentences, the ones containing the query term are passed onto the DPF. A query term may consist of more than one word.

2.3 Main components

The main components of the DPF system consist of the key phrase pattern matcher, word counter, and sentence locator. These components process the candidate sentences separately, and each produces its own score for a sentence. A final score computed from the component scores is used for ranking the sentences. Each component is now described.

Key Phrase Pattern Matcher

The key phrase pattern matcher attempts to detect descriptive phrases of a query noun by simple pattern matching using lexically motivated text fragments. The key phrase patterns are listed with examples in Table 1.

Table 1: Key phrase pattern matcher

Type	Pattern	Example
such as	(*dp** **such** | **such** *dp*) **as** *qn**	... injuries **such as** break bones ...
and other	*qn* (**and** | **or**) **other** *dp*	... break bones **and other** injuries ...
especially	*qn* (**especially** | **including**) *dp*	... injuries **especially** bread bones ...
is a	*qn* (**is** | **was** | **are** | **were**) (**a** | **an** | **the**) *dp*	... Tony Blair **is a** politician ...
Acronym	*qn* **(***dp***)** or *dp* **(***qn***)**	... MP **(**Member of Parliament**)** ...
qn, dp,	*qn***,** (**a** | **an** | **the**) *dp***,**	... Tony Blair**, the** politician**,** ...

*A *dp* and *qn* denote a descriptive phrase and query noun, respectively.

The first three types are derived from Hearst's investigation to identify hypernym relations for improving the WordNet thesaurus (Hearst, 1998). The fourth type is probably the most intuitive pattern, detecting descriptive phrases by the IS-A relation. Resolving acronyms is also attempted (the fifth type). The last type is called a comma parenthesised apposition. This pattern locates appositive phrases of a query noun, which are often used to describe the noun in more detail (Coates-Stephens, 1993). The following variants of this pattern were also devised in the DPF system.

- *qn*, **which** (**is** | **was** | **are** | **were**) *dp*,
- *qn*, (**a** | **an** | **the**) *dp* (**.** | **!** | **?**)
- *qn*, *dp*, (**is** | **was** | **are** | **were**)

It should be noted that these patterns using the text fragments and other punctuation are simple but effective at locating the corresponding descriptive phrases of a given query noun. In other words, when a query noun and these patterns co-occur, descriptive phrases are likely to be found in the sentence.

The sentence score produced by this component depends on the previously observed accuracy of the pattern matched. The accuracy was investigated prior to the first experiment and is described in the tuning section (2.4).

Word Counter

When building the DPF, it was anticipated that a descriptive phrase or its component words might commonly co-occur with the query across a range of sentences and that the presence of such multiple occurrences might be exploitable. Therefore, the frequency of occurrence of the sentence words (excluding stopwords) was computed across the set of matching sentences and the 20 most frequent were recorded. It was assumed that most of these common words would be component words of the description. When scoring a sentence, each was examined for the presence of the common words and assigned a score based on the sum of the word frequencies. As a result, the sentences containing more of the common terms were given a higher score. As reported in Joho and Sanderson (2000), exploiting repetition of descriptions was tested and found

to be successful in the preliminary DPF work. More recently, aspects of this approach are to be found in QA systems utilising redundancy (Clarke *et al.*, 2001).

Sentence Locator

The sentence locator component provides a candidate sentence with a score based on the location of the sentence in a document. It seemed reasonable to expect that if a noun was used a number of times within a document, then any accompanying description of it was to be found nearer the start than the end. Therefore, the ordinal position of sentences containing the query noun (e.g. 1st 2nd, 3rd, etc) is noted with earlier sentences given a higher score.

Final Score

The final score for each candidate sentence is given by a simple linear combination:

$$FinalScore = aKP + bWC + c(d - SL)$$

where KP is the weighted score from the key phrase pattern matcher, WC is from the word counter, and SL is from the sentence locator. The values of a, b, c, d are the tuning constants, and set to 2000, 1, 75, and 500 respectively after tuning (described next). The large number of the first constant, a, ensures that those sentences matching any pattern are ranked higher than those unmatched. The system ranks the candidate sentences based on the final score.

2.4 Tuning the system

As mentioned above, it was necessary to tune the system to obtain an optimal performance from the key phrase pattern matcher and combination of the final score. Therefore, a descriptive phrase test collection was created from LA Times TREC Collection articles (1989 & 1990, 475MB). By dividing it into two: half (the training set) was used for tuning the system, and the other half (the test set) was used later for evaluation. Details of queries and relevance judgements are to be found in the experimentation section.

2.4.1 Accuracy of key phrase pattern matcher

While the key phrase pattern matcher was implemented, it was clear that some of the patterns were going to perform better to locate descriptive phrases than others. Therefore, the relevance of sentences extracted from the collection was assessed to measure the accuracy of the patterns. The result is shown in Table 2 along with their coverage.

Table 2: Accuracy of key phrase pattern matcher

	Code	Not relevant	Relevant	Total	Accuracy
No pattern	na	6424	872	7296	12.00%
especially	es	0	0	0	00.00%
qn, dp,	ap	89	63	152	41.40%
is a	ia	23	18	41	43.90%
including	in	20	17	37	45.90%
or other	oa	1	1	2	50.00%
such as	sa	59	59	118	50.00%
acronym	ac	14	23	37	62.20%
and other	ao	9	23	32	71.90%

As can be seen, all the patterns are relatively rare (compare with numbers for no patterns) which confirms the need to search large corpora. Among the patterns, ”and other” was proved to be most accurate. These figures were used in the ranking of sentences with those containing the more accurate key phrases getting a higher score.

2.4.2 Combination for the final score

A series of tests were run on different combinations of the scores gained from the main components. A weighted sum of the scores which was found to work best was as shown in the previous section.

A sample output of the DPF system used in the third experiment is presented in Table 3. From the left side, each column shows a rank, final score, pattern matched (c.f. the 2nd column of Table 2), score corresponding to the pattern, score from the word counter, score from the sentence locator, and a candidate sentence, respectively. In the example, the query term *Tony*

Blair was highlighted to help users to locate a descriptive phrase.

3 Experiments

Three experiments were carried out to evaluate the effectiveness of the DPF. Details of each experiment are described in the following subsections, but we first briefly summarise the direction taken across the experiments.

The first experiment was aimed simply at evaluating the effectiveness of our approach to descriptive phrase finding (Joho and Sanderson, 2000). With the success of the it, the second experiment was aimed to evaluate the effectiveness of the DPF system in a much larger collection, the Web (Joho *et al.*, 2001). This experiment introduced a more stringent relevance judgement of descriptive phrases. The third experiment (the main focus of this paper) examined the impact of using different search engines as ”backends” to the DPF system. Here the effect of varying collection coverage and search engine ranking techniques was explored.

3.1 Experiment 1

The first experiment focused on two aspects of the effectiveness of the DPF system. The first was a full evaluation of the system, the second an initial investigation of the impact of collection size on accuracy.

3.1.1 Method

Although documents were taken from the TREC LA Times test collection, none of the queries of that collection were found to be appropriate for testing a DPF. Consequently, a set of 50 queries was created by the authors for evaluation. Because of the time frame of the documents being searched (1989-90), queries were chosen for which a description was expected to be found. A total of 16,111 candidate sentences was found for the 50 queries and was evaluated for relevance by the experimenters. Relevance was defined (leniently) as a descriptive phrase that contained some information to help a user understand more about the query noun they queried on. It should be noted that judging relevance was on sentences and not on extracted descriptions or other units

Table 3: Sample output of the DPF system

Rank	Score	Code	KPW	Word	Sentence	Sentence
1	56998	ao	9.824	0	4	**Tony Blair**, Gordon Brown, Robin Cook and other members of the Cabinet, who have their Offices in the
2	55460	ap	8.794	522	2	**Tony Blair**, leader of the Labor Party and Prime Minister of Britain, was born on May 6, 1953, at 6.05 in the morning in Edinburgh,
3	55396	ap	8.794	383	8	HOWEVER, **Tony Blair**, the UK prime minister, on Tuesday put pressure on the Royal Bank of Scotland to continue its support for Huntingdon Life Sciences, the drug testing laboratory that is being targeted by animal rights activists.
4	55379	ap	8.794	441	9	To our astonishment, the BBC led off its news program not with the O. J. Simpson verdict but with the keynote speech that **Tony Blair**, the Labour leader, had given earlier that day.
5	55321	sa	8.86	326	4	The modern open mouthed smile exposing the teeth is a particular favourite of politicians such as Prime Minister **Tony Blair**.
6	55280	ap	8.794	417	8	I was busy in my flat putting small pieces of fruitcake in jam jars, when all of a sudden the front door burst open and in walks **Tony Blair**, the Labour front-bencher widely tipped by the Tory media to fill the leadership gap left by the death of John Smith.
7	55160	ap	8.794	297	6	Yet for someone in full command of his party and his job, **Tony Blair**, the prime minister of Britain, remains a surprisingly unknown figure.
8	55109	ia	8.75	259	2	**Tony Blair** is the current Prime Minister of the United Kingdom.
9	55085	ap	8.794	297	10	Yet for someone in full command of his party and his job,**Tony Blair**, the prime minister of Britain, remains a surprisingly unknown figure.
10	54925	ia	8.75	0	1	**Tony Blair** is an Alien!

of text smaller than a sentence. This made it possible to automatically process the results of the DPF system in a similar manner to traditional IR test collections.

Evaluation concentrated on rank-based measures: precision at ranks 1, 5, and 10. In addition, the percentage of successfully answered queries was also calculated. This was the percentage of queries for which at least one correct answer was found within a specified rank position.

3.1.2 Results

The results were as follows.

Collection size

In this experiment random samples of the test set were taken and the effectiveness of the key phrase pattern matcher was measured for each of these samples. Results from this experiment are shown in Table 4. Samples taken were for 10%, 25%, 50%, 75% and 100% of the test set.

Table 4: Precision (Key Phrases) and Collection size

P. at rank	100%	75%	50%	25%	10%
1	0.75	0.78	0.69	0.63	0.62
5	0.51	0.51	0.46	0.38	0.32
10	0.42	0.4	0.35	0.28	0.24

The results of this test show that as the collection got smaller, the effectiveness of the system reduced due to the decreasing likelihood of the system finding a sentence holding one of the key phrases. This result echoes that obtained by Hawking on the VLC collection described above.

Precision and recall

The effectiveness of each of the individual sentence ranking criteria was tested along with combination formula derived above. Unlike most IR test collections, the ratio of relevant to non-relevant was relatively high. Therefore, it was important to establish a random baseline as well. To achieve this, for each of the fifty queries, 100 random orderings of the sentence collection (in the test set) were generated and the average effectiveness of these cumulative runs was measured.

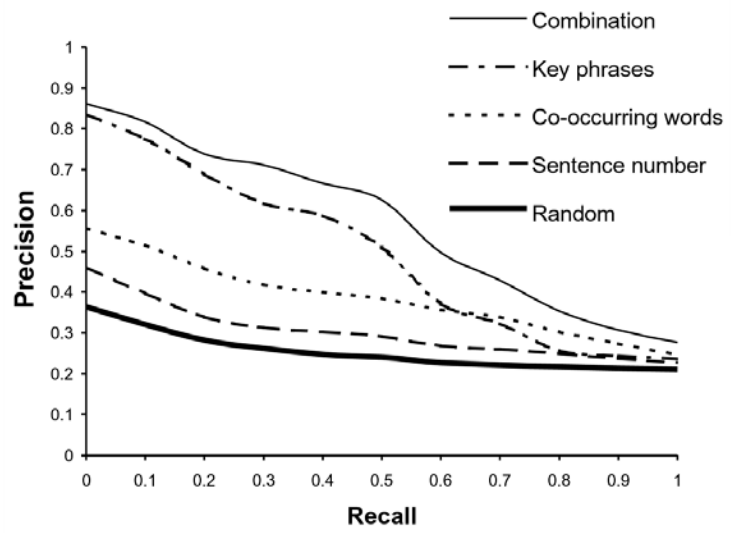


Figure 2: Recall-precision graph of four strategies

A recall-precision graph was plotted showing the effectiveness of the four strategies plus random retrieval (see Figure 2). As can be seen, the three criteria and their combination do better than random retrieval. A sentence ranking based purely on the key phrase weights was extremely effective, except for high recall situations where the co-occurring word counting method was better. The most effective technique for finding descriptive phrases, however, was the combination formula, which, through a t-test, was found to be significantly better than any of the other methods, including key phrases. The difference between the combination and key phrase methods was found to be significant at $p < 0.05$ for recall levels 0.1 and 0.2 and significant at $p < 0.01$ for all higher values of recall. When evaluated with precision oriented measures a similar picture emerged.

Table 5 shows precision measured at rank positions ranging from one to one hundred. As can be seen, the combination method is consistently better than the key phrase. Like the recall-precision graph, significance testing was performed: the combination method was found to be significantly better than key phrase for all rank positions from five through to 100 ($p < 0.01$). The percentage of queries with at least one correct answer in the top n was also calculated. Here n was chosen to be 5 and 10 as it was felt that a user would be willing to look through this number

Table 5: Precision of four strategies

Precision at rank	Combination	Key phrase	Co-occur. Words	Sentence Number	Random
1	0.76	0.75	0.37	0.25	0.20
5	0.57	0.51	0.35	0.27	0.20
10	0.46	0.42	0.35	0.27	0.20
15	0.42	0.36	0.33	0.24	0.20
20	0.38	0.32	0.32	0.23	0.20
30	0.32	0.26	0.28	0.22	0.19
100	0.17	0.15	0.16	0.15	0.14

of sentences. For the best performing method (combination) 90% of the queries had a correct answer in the top 5 (compared with 22% for random) and 94% correct in the top 10 (c.f. random 31%)

3.2 Experiment 2

The results of the first experiment proved the feasibility of our simple approach to locate descriptive phrases from free-text. It was speculated that the success was due to a relatively large amount of text searched. Subsequently, the second experiment was designed to evaluate the effectiveness of the DPF system with an even larger collection, the web.

3.2.1 Method

The web was chosen as the collection of the second experiment and this was achieved by using Google[1] as the IR component. The collection of Google consists of approximately 1 billion documents, four orders of magnitude larger than the 100,000 document LA Times collection. The retrieved web documents were fetched from a maximum of 600 URLs returned by Google in response to each query.

The criterion of relevance judgement of descriptive phrases was revised for the second experiment, and a strict definition of relevance was devised as the lenient relevance of the previous

[1] `http://www.google.com`

experiment was felt to be unrepresentative of the DPF task. Key answers were created for each query: defining clearly the descriptive phrase expected. For example, “the Prime Minister of Great Britain” was the key answer for the query *Tony Blair*. The sentence ‘Tony Blair is the current Prime Minister of the United Kingdom’ was regarded as relevant, but ‘Tony Blair is a political leader’ was not. This relevance will be referred as *strict relevance* while the existing method will be referred as *lenient relevance* in this paper.

A total of 50 queries were tested with the strict test using key answers. Unlike the previous experiment, it was not possible to judge relevance of all matching sentences due to collection size. Consequently, relevance judgements were made on the top 20 sentences ranked by the DPF system for each query.

3.2.2 Results

Stringent Test

The percentages of successfully answered queries in the top 5, 10 and 20 sentences using 50 queries were measured with strict relevance. Despite of the stringent relevance criteria, it was shown that the DPF system retrieved at least one relevant descriptive phrase for 66%, 82%, and 88% of the queries in the top 5, 10, and 20 sentences respectively.

Lenient Test

As a means of comparison, a repeat of the experiment 1 test was also conducted using lenient relevance definitions and the same set of 50 queries as used for the LA Times collection. In addition 46 new queries were devised, thus, a total of 96 queries were used for the web collection. The comparison with the result of the first experiment is illustrated in Table 6. This shows that a significant improvement over the previous experiment was achieved by moving to a larger collection.

These results show that the DPF searching the Web works dramatically better than the previous experiment using LA Times. As was shown in previous work, the size of the collection impacts on the effectiveness of the system. This is because by searching a larger collection, there is a better chance of locating a relevant descriptive phrase in the format of one of the searched

Table 6: Percentages of successfully answered queries: Comparison with the previous experiment (Lenient relevance)

	LA Times (N=50)	Web (N=96)	Improvement (%)
Top 5	90%	100%	+11.11
Top 10	94%	100%	+6.38

for key phrases. However in the previous work, there appeared to be an upper bound on the accuracy of the descriptive phrases alone. By searching a much larger collection it is speculated that the cross document term occurrence statistics used contributed significantly to improving the effectiveness of the system.

3.3 Experiment 3

At the end of experiment 2, it was also speculated that Google's PageRank system (Brin and Page, 1998) may have impacted on the effectiveness of the DPF system. Each Web search engine has different collection sizes and coverage in addition to different ranking techniques. Therefore, in the third experiment, our attention was turned to the effects of using different search engines as the backend of the DPF system.

The size of retrieved documents provided to the DPF was also examined in this experiment. Previously, the contents of the top 600 URLs returned for each query was used, but this value was never tested. Consequently, it was decided to investigate to what extent different numbers of retrieved URLs affect the performance of the system.

3.3.1 Method

The choice of search engines for this experiment was not so straightforward since the information about the internal mechanisms of each is hard to come by. As a result of informal discussions, three search engines with very different ranking techniques were selected: Google, WebTop, and DirectHit. Google was used again as it had the largest collection and used the PageRank

algorithm. WebTop[2] was chosen because the ranking method is believed to be based on BM25 ranking (e.g. Robertson *et al.*, 1995). DirectHit[3] was chosen because its ranking method incorporates popularity information making it different from the other two.

To examine the effect of different numbers of URLs, samples were taken from the documents fetched from the top 100, 210 and 600 URLs for each query. The second size (i.e. 210) was chosen because this is DirectHit's maximum number of URLs one can retrieve.

To investigate the effect of ranking techniques on the DPF system, it was necessary to examine how each search engine ranked a common set of web pages for each query. It was found that due to DirectHit's 210 limit on pages returned it was not usable in this experiment. Therefore, intersection sets from WebTop and Google only were created by retrieving 1,000 URLs per engine, noting the URLs in common and their rank position in both systems. Then the top 55 URLs as ranked by Google and the top 55 as ranked by WebTop were extracted. Since the URLs in the common set are identical between two search engines but ranked in different positions, the intersection set enables us to focus on the ranking techniques of two search engines and ignore collection coverage. This approach was problematic however, as many queries did not yield a large intersection set. Hence, only queries that had more than 80 common URLs were used for this analysis.

The same scheme of relevance judgement as the previous experiment, namely the strict judgements, were used in this experiment. The same 50 queries from experiment 2 were used. The measures of evaluation were the percentage of successfully answered queries, and precision based on the number of relevant sentences in the top 20 for each query. Because of the lack of intersections, only 28 of the 50 queries were used in the rank technique test.

3.3.2 Results

Effect: Number of URLs

The first test focused on different numbers of URLs. More URLs will increase the chances of locating more descriptive phrases, but at the risk of introducing noise. Samples of 100, 210 and

[2] `http://www.webtop.com`

[3] `http://www.directhit.com`

600 URLs were taken from the search engines. Figure 3 shows the percentages of successfully answered queries in the top 20 sentences.

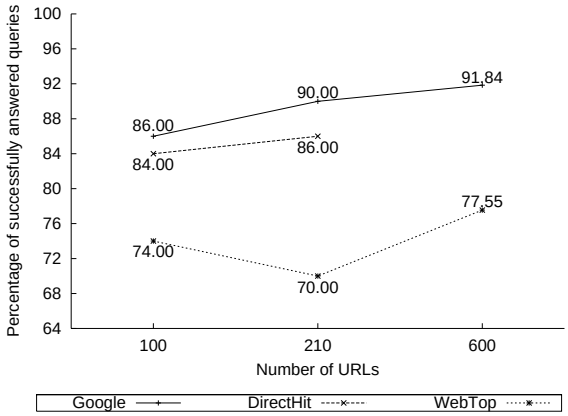


Figure 3: Percentages of successfully answererd queries: At least one relevant descriptive phrase in the top 20 sentences (*Strict* relevance).

Generally, the performance of the system increased with the number of URLs except at the 210 URLs from WebTop. It would appear that any concerns about introducing noise from excessive retrieved documents were not founded.

Table 7 shows the precision for the top 20 sentences retrieved by the system. In most cases, precision was increased when more URLs were used. This echoes the previous result.

Table 7: Precision for the top 20 sentences

No. of URLs	Google	WebTop	DirectHit
100	0.4927	0.4478	0.5384
210	0.5141	0.5110	0.5453
600	0.5251	0.5209	—

Effect: Ranking method

The effect of the ranking method in different search engines was also examined using intersection

sets of URLs from Google and WebTop. The intersections were generated from 28 queries and consist of the top 55 ranked URLs from each search engine.

The bottom of Table 8 shows the percentages of successfully answered queries along with precision for this test. WebTop's sets performed as well as Google's at the top 20 sentences on all the levels. However, the precision of Google's sets was higher than WebTop's. It is speculated that this similar performance between two search engines is due to the small number of URLs in the intersection sets.

From the result, it is possible that different ranking methods also affect the performance of the DPF system. It would appear that Google's PageRank system provided more chances for identifying descriptive phrases than WebTop. However, a greater overlapping collection is needed to analyse this effect further. The comprehensive results of the third experiment including the performance for the top 5 and 10 sentences of all tests described above are shown in Table 8.

Table 8: Overall results of successfully answered queries (%) and precision

No. of URLs	Search Engine	Top 5	Top 10	Top 20	Prec.
100	Google	72.00	78.00	86.00	0.4927
	WebTop	52.00	64.00	74.00	0.4478
	DirectHit	72.00	82.00	84.00	0.5384
210	Google	76.00	80.00	90.00	0.5141
	WebTop	58.00	66.00	70.00	0.5110
	DirectHit	74.00	82.00	86.00	0.5453
600	Google	80.00	82.00	92.00	0.5251
	WebTop	63.27	73.47	77.55	0.5209
	DirectHit	——	——	——	——
55 (Intersec.)	Google	82.14	85.71	92.86	0.5960
	WebTop	82.14	85.71	92.86	0.5814

4 Related Work

Fujii and Ishikawa (2000, 2001) have developed a system that has a similar aim to the DPF system. Their system aims to extract *encyclopedic knowledge* of technical terms from the Web. It is anticipated that those extracted descriptions can be applied to answer a type of questions

such as “What is X?”. Their system also used pattern matching based on the *Is-a* relations for initial detection of candidate descriptions from the documents retrieved by Google. Resolving particular forms of HTML tags (e.g. <DD>, <DT>, and <UL>) were also applied. The candidate descriptions were then clustered into pre-defined 19 domains based on a variant of the Bayesian probabilistic model, which was trained by a dictionary used in a commercial machine translation system. In each cluster, descriptions were ranked based on an N-gram probability. Three top ranked descriptions from each domain were the output of the system.

To compare with their system, the DPF system was implemented by a simpler approach without relying on NLP-based techniques. They reported that their system retrieved at least one relevant descriptions in the top 3 for 90% of 85 queries. Our system retrieved at least one relevance descriptions in the top 5 sentences for all the 96 queries as described in the second experiment. From an informal comparison of the performance, it seems our simpler approach can achieve as good as or better performance than their system. In addition the DPF system is free from any domain restriction.

5 Conclusion

In this paper we presented a system that retrieved descriptive information about query nouns, and described three large scale experiments. The design of the system was motivated by findings from the VLC track of TREC, that is, a larger collection provides a better chance to retrieve relevant information, thus, a simple approach can achieve a high performance without complicated techniques.

This paper also described two criteria for relevant judgement of retrieved descriptive phrases: lenient and strict. The latter stringent relevance criterion using the key answers was used to evaluate the effectiveness of the system to retrieve highly relevant descriptive phrases. In lenient relevance, the system retrieved at least one relevant description in the top 5 sentences for 100% of the tested queries, while the strict test showed that the system retrieved at least one highly relevant description in the top 20 sentences for 92% of the queries. These results justified our approach.

Effects of different search engines were also investigated. Three search engines were used to retrieve the documents that the system searched for descriptive phrases. Both the number of documents and ranking techniques were found to effect the performance of the system. There seemed to appear more noises in a larger number of URLs to be used, but the system could locate relevance descriptions among them. The effect of ranking techniques was found to be less significant than the number of URLs. Overall, a set of documents retrieved by Google provided the best chance for the system to locate more relevance descriptive phrases.

Future work may be focused on user ability to locate descriptive phrases in the candidate sentences. In our experiment only query terms were highlighted to help users to spot a descriptive phrase. However, a more sophisticated interface design would improve the effectiveness of such a task.

Acknowledgments

This work was carried out as a part of the CiQuest project, which is 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

Brin, S. and Page, L. (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine. In: *Proceedings of the 7th International World Wide Web Conference (WWW7)*, Brisbane, Australia. Available from `http://www7.scu.edu.au/programme/fullpapers/1921/com1921.html` [Accessed: 27/09/01].

Clarke, C. L. A., Cormack, G. V., and Lynarn, T. R. (2001). Exploiting Redundancy in Question Answering. In: *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 358–365, New Orleans, LA. ACM.

Coates-Stephens, S. (1993). The Analysis and Acquisition of Proper Names for the Understanding of Free Text. *Computers and the Humanities*, **26**, 441–456.

Fujii, A. and Ishikawa, T. (2000). Utilizing the World Wide Web as an Encyclopedia: Extracting Term Description from Semi-Structured Texts. In: *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL-2000)*, pp. 488–495, Hong Kong.

Fujii, A. and Ishikawa, T. (2001). Organizing Encyclopedic Knowledge based on the Web and its Application to Question Answering. In: *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL-EACL 2001)*, Toulouse, France. *To appear*.

Hawking, D. and Thistlewaite, P. (1997). Overview of TREC-6 Very Large Collection Track. In: Voorhees, E. M. and Harman, D. K. (eds), *Proceedings of the 6th Text REtrieval Conference (TREC-6)*, pp. 93–106, Gaithersburg, MD. NIST.

Hearst, M. A. (1998). Automated Discovery of WordNet Relations. In: Fellbaum, C. (Ed.), *WordNet: An Electronic Lexical Database and Some of its Applications*, pp. 131–151. MIT Press, Cambridge, MA.

Joho, H. and Sanderson, M. (2000). Retrieving Descriptive Phrases from Large Amounts of Free Text. In: Agah, A., Callan, J., and Rundensteiner, E. (eds), *Proceedings of the 9th International Conference on Information and Knowledge Management (CIKM)*, pp. 180–186, McLean, VA. ACM.

Joho, H., Liu, Y. K., and Sanderson, M. (2001). Large scale testing of a descriptive phrase finder. In: Allen, J. (Ed.), *Proceedings of the 1st International Conference on Human Language Technology Research (HLT 2001)*, pp. 219–221, San Diego, CA. Morgan Kaufmann.

Kupiec, J. (1993). MURAX: A Robust Linguistic Approach For Question Answering Using An On-Line Encyclopedia. In: Korfhage, R., Rasmussen, E. M., and Willett, P. (eds), *Proceedings of the 16th Annual ACM SIGIR conference on Research and Development in Information Retrieve*, pp. 181–190, Pittsgurgh, PA. ACM.

Radev, D. R. and McKeown, K. R. (1997). Building a Generation Knowledge Source using Internet-Accessible Newswire. In: *Proceedings of the 5th Conference on Applied Natural Language Processing*, pp. 221–228, Washington, DC. ACL.

Robertson, S., Walker, S., and Hancock-Beaulieu, M. (1995). Large Test Collection Experiments on An Operational, Interactive System: Okapi at Trec. *Information Processing & Management*, **31**(3), 345–360.

Srihari, R. and Li, W. (1999). Information Extraction Supported Question Answering. In: Voorhees, E. M.

and Harman, D. K. (eds), *Proceedings of the 8th Text Retrieval Conference (TREC-8)*, pp. 185–196, Gaithersburg, MD. NIST.

Appendix 4

Joho, H, Sanderson, M and Beaulieu, M. Large scale testing of document frequency's relatedness to term specificity. To be submitted to *Information Retrieval*.

Large Scale Testing of Document Frequency's Relatedness to Term Specificity

Hideo Joho, Mark Sanderson, and Micheline Beaulieu*

Abstract

It has been long understood that the frequency of occurrence of a term in a corpus is correlated with the term's specificity. The less frequently a term occurs, the more semantically specific it tends to be. It has also been long known that such a correlation is not perfect. A small-scale study of the relationship between occurrence and specificity was conducted examining the occurrence frequencies of a few hundred words in a corpus of tens of thousands of documents. The work in this paper goes beyond the past study by examining tens of thousands of terms and by using a Web search engine to determine their frequency from a corpus that is larger by four to five orders of magnitude. The study provides a greater understanding of the relationship of frequency to specificity, and through an examination of term co-occurrence, describes the impact of the results on the building of ontologies/topic hierarchies. In addition, results and views on the quality of the Web as a corpus, compared to other more classic corpora, are presented.

*Address: Department of Information Studies, University of Sheffield, Western Bank, S10 2TN, UK. Email: {h.joho, m.sanderson, m.beaulieu}@sheffield.ac.uk.

1 Introduction

Topic hierarchies have long interested researchers in information retrieval (IR), computational linguistics (CL), and other related areas. In IR, the structures have been employed to aid users in browsing sets of documents and in helping them formulate or later expand their queries. In CL, such hierarchies have been used as a resource for other language-based tasks.

The means of automatically creating such hierarchies remains an active subject, where the nodes of the hierarchy (composed of concepts or individual words) are arranged in some taxonomic structure with general concepts at the top of the hierarchy leading to related and more specific concepts below. Methods for locating words or phrases that would be good candidate concepts and means of determining their relationship (through some measure of co-occurrence) have been well studied (Grefenstette, 1992; Anick and Tipirneni, 1999). Somewhat less examined, however, is the issue of term specificity: given a pair of terms/concepts that have been found to be related, how does one determine which is a more specific concept?

To this end, we have examined one of the basic properties of terms: document frequency. The notion of determining term specificity through document frequency is not new. Salton suggested such an approach to ordering terms in a hierarchy in his 1968 book (Salton, 1968). However, it was suggested soon after that document frequency and specificity may not correlate strongly: Sparck Jones (1972), in her paper introducing *inverse document frequency* (IDF), implied that the two were not the same thing. However, the focus of her paper was not on specificity and no testing of the relationship between the two was conducted.

Salton's idea of using frequency was found again in work by Forsyth and Rada (1986) where a limited scale concept hierarchy was constructed and related terms were ordered by frequency. It would appear, however, that throughout this early work, little actual testing of the relationship between frequency and specificity was conducted.

More recently, a small scale test was conducted by Caraballo and Charniak (1999) who examined a fragment (200 words) of the WordNet hypernym hierarchy measuring the frequency of

occurrence of words in the hierarchy from a corpus (1987 Wall Street Journal). They showed that a correlation did exist. However, the size of corpus examined and the range of words tested (both in terms of number and type[1]) was limited.

The work in this paper constitutes a significant expansion of the experiment conducted by Caraballo in terms of corpus size and number of terms examined. An alternative approach to dealing with word ambiguity is incorporated into the experimental design as well. The paper presents a large-scale experiment that examined document frequency on the Web (a corpus many times larger than WSJ) and its relatedness to term specificity. It also attempts to assess the quality of the Web as a corpus within the context of determining specificity.

The rest of the paper starts with a review of existing work in this area. Next, the design and results of an experiment examining specificity in document collections is outlined, followed by a second experiment focusing on sets of retrieved documents. The implications of the results for the development of topical hierarchies are then discussed before the paper concludes.

2 Related Work

As mentioned above, large scale studies that examined the relationship between document frequency and term specificity are few. One such work, that indicates word counting-based measures are able to determine specificity, is Weinberg and Cunningham (1985) who carried out a test to examine the relationship between terms in MeSH and the number of documents in MEDLINE (1966-1985). They selected a hierarchical tree under the term *Endocrine Diseases*, a central topic in MeSH that contained about 100 terms in four hierarchical levels. They also selected another similar size of tree under the term *Environment* as a peripheral topic. They found a negative correlation between the depth (level) of the hierarchies and number of documents in which terms occurred. The negative correlation of the terms in the central tree (-0.20422) was

[1]It appears that many of the words used in the study had one sense only. This, it would seem, was done to avoid the problems of correlating corpus-based word frequency with the frequency of specific senses of words

larger than in the peripheral one (-0.13045), although both correlations appeared to be weak. This provides some evidence of a relationship, but the small sample size and limited domain of the documents and thesaurus implied that the study may not generalise well.

More often, researchers have attempted to interpret term specificity through a statistical measure (Robertson, 1974). As mentioned before, Sparck Jones (1972) suggested that specificity should be measured by the frequency of occurrence of a term, where a less frequent term was regarded as more specific. She commented that this type of specificity was not necessarily to be the same as semantic one but it was useful for retrieval systems. In a similar context, Barker, Veal, and Wyatt (1972) estimated term specificity by determining the total number of documents containing a term, and calculating what proportion was relevant. This was designed to determine how specific a term was to a particular query.

Document frequency was used to determine term specificity along with co-occurrence information in Sanderson and Croft (1999). In their approach a topic hierarchy[2] was derived from a set of retrieved documents by a process known as subsumption. The subsumption algorithm attempted to identify which term of a pair was more general (or specific), by examining sets of documents in which both terms occurred. More specifically, if a set of documents in which term y occurred was a large subset of the documents in which term x occurred, then term y was said to be *subsumed* by term x. Since term y occurred fewer times than term x, and since y's document list was a part of x's, term y was judged to be more specific than term x. In other words, their approach attempted to identify related terms by co-occurrence information, and determine which was more specific by document frequency.

As with many previous works, no test was conducted to examine the correlation between term specificity and document frequency. However, a user-based study provided limited evidence of the ability for document frequency to order terms based on specificity. However, since their method combined the evidence from document frequency and co-occurrence information, the performance based solely on document frequency was still not clear.

[2] which was referred to a *concept hierarchy* in their paper.

As can be seen, researchers have been using document frequency as a means of determining term specificity, but perhaps surprisingly a large scale test has not been carried out.

With regards to the means of determining term specificity, recent natural language processing (NLP) based approaches (Grefenstette, 1994; Woods, 1997; Anick and Vaithyanathan, 1997) are likely to be more precise than word counting for certain word types. However, comparing these linguistically motivated techniques, document frequency seems to be superior in handling a larger range of words, which can be crucial in many applications. Generating document frequency lists from a large corpus is computationally more efficient than parsing. More precise measurements based on linguistics can then co-operate with the limited cases where the simple counting fails. Therefore we believe that it is important to investigate and understand the ability of document frequency to determine specificity, so that an integrated technique can successfully benefit from both approaches.

Our study centres on the Web as a corpus for determining semantic information. Although it is a very large corpus, size alone should not be the reason for using it, therefore, in addition to the study of specificity, a comparison of the Web corpus and smaller more commonly used corpora is also conducted. To the best of our knowledge, no such comparative work has been conducted before[3].

3 Experimental Design

The aim of our experimental work was to test on a large-scale, the ability of document frequency to determine specificity. The experimental design has its roots in Caraballo and Charniak's approach of measuring the accuracy of document frequency by comparing the measure's success in ordering word pairs taken from WordNet. The data we used in our experiment were as follows. Approximately 45,000 noun words and phrases in WordNet[4] (Miller, 1990) were used

[3] A possible exception might be Kilgarriff (2001) who explored a means of measuring similarity between corpora. His work, however, did not use the Web but small tagged corpora.

[4] The version was 1.6.

Figure 1: Sample query in Google (Submitting a term *eye contact* with quotations. The estimated number of URLs containing the term was 368,000 and this would be the document frequency of *eye contact*.)

for our experiments. Document frequency was determined using Google[5] (at the time of experiments over a billion Web documents said to be in its collection) by recording the number of documents retrieved in response to each WordNet word (or phrase) submitted as a query (see Figure 1). Document frequency for a synset was estimated by averaging the document frequency of each member term.

3.1 Word Sense Disambiguation

As with almost any study involving the determination from corpora of a term's attributes, the problem of word sense was considered to determine the viability of our experimental approach. The essential problem was that the WordNet hypernym chains are composed of word senses and we wished to know their frequency of occurrence. Of course, the corpus we were using (the Web) was not sense tagged. In order to estimate frequency of occurrence of senses from a corpus of words, it was necessary to focus our study on a sub-set of senses that one could be more confident about measuring.

Research from Sanderson and van Rijsbergen (1999) showed that the commonest sense of a word often accounted for the outright majority of the word's occurrences in a corpus. If one focused the specificity study on only a term's commonest sense, one could assume that the frequency of occurrence of the term in a corpus was reasonably well correlated to the occurrence of its pre-

[5]http://www.google.com/

vailing sense. With such an approach, there remains the problem of determining the commonest sense of a term. Given the number we wish to study (tens of thousands), the commonest sense needed to be found automatically. Therefore, we chose to use the commonest sense defined by WordNet.

The means that the creators of WordNet used to determine the frequency of occurrence of a word's sense was two fold: first if the word occurred in the Sense Eval corpus[6], the commonest sense was measured from there; if the word did not occur in that corpus, then it is our understanding that the commonest sense was determined by the WordNet creators based on their lexical/world knowledge.

At the core of our experimental design therefore, was the assumption that the commonest sense of a term accounts for the great majority of that term's occurrences, and that the definition of commonest sense in WordNet was correct in our corpus. The later assumption is perhaps the one that should be questioned the most, after all, what if the corpus one is examining focuses on a different domain of texts from that used in WordNet? Words in that domain are likely to have different commonest senses. Again here, an assumption was made that sense usage in the Web would correspond to that presumed by the creators of WordNet. Although it may appear to some that this is one assumption too far, experimental results presented in Section (4.1) appear to support it and therefore support the experimental design.

3.2 Hypernym chain and synset

Among the semantic relations defined in WordNet, *hypernymy* (superordinate) orders words by their specificity. Terms that have no hyponyms can be regarded as one of the most specific terms in WordNet. A maximal hypernym chain of such a term can be obtained by the iteration of tracing the direct hypernym term to the top of the hierarchy. An example maximal hypernym chain of the term *eye contact* is 'eye contact' $>$ 'contact' $>$ 'interaction' $>$ 'action' $>$ 'act, human action, human activity'.

[6] a subset of the Brown corpus, where the senses of its constituent words were manually disambiguated

Table 1: Chain length and the number of chains

Chain length	Original	Disambiguated
2	38	−−
3	490	157
4	2,224	733
5	6,046	2,696
6	13,792	5,297
7	12,681	5,681
8	10,024	4,793
9	7,369	2,773
10	3,917	1,480
11	1,945	863
12	718	368
13	424	262
14	203	115
15	48	24
16	1	0
Total	59,920	25,242

A total of 59,920 maximal hypernym chains (referred as simply *hypernym chains*, or *chains* from now on) were found in WordNet and the length of chains ranged from 2 to 16. The chains in length 2 were removed since they were too short for our experiments. Due to the disambiguation issue described in the previous subsection, 25,424 hypernym chains consisting of only primary sense synsets were used. Table 1 shows the number of chains according to the length in both the original WordNet and disambiguated setting.

As can be seen, the disambiguated set of hypernym chains was approximately 40% of the original chains. The three of the most frequent length in the disambiguated set were six, seven, and eight. These disambiguated hypernym chains were used as the basic data in our experiments, which are described in the next section.

4 Experiments

A series of heuristic experiments was carried out to examine the relationship between taxonomic term specificity and document frequency. The experiments consisted of three parts: the

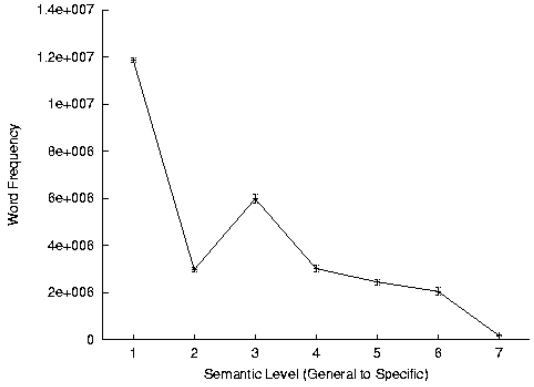


Figure 2: Average document frequency (Length 7)

first analysed the average document frequency of each level of hypernym chains; the second studied probability of parent synsets holding a higher document frequency than the next lower child synsets; and finally the third observed the effects of co-occurrence information on such probability.

4.1 Average document frequency

The first part of our experiment the average document frequency of synsets was calculated for each level of a set of hypernym chains that were the same length. To illustrate, the averages of length 7 chains are plotted in Figure 2.

The graph shows that the most general level (1) has the highest average document frequency and the most specific level (7) has the lowest, respectively. However a monotonically decreasing line between them was not found. Level 3 has the second highest average. Similar results were found in the other sets except length 3 and 4. The graphs for the other sets of length are shown in Figure 3.

It is not entirely clear why a particular middle level has higher document frequency, but it is speculated that the terms at such a level belong to a *basic level category* (Lakoff, 1987), the most common level of detail in categorisation. In learning process of vocabulary, Lakoff explains that

there is a category that one learns first. An example of an animal basic level category is bird, a word that a child might learn first; other levels, e.g. more specific such as dove, or more general such as animal, are learnt later. Because of their familiarity, terms in basic level categories are likely to have higher document frequency than those below or above.

As can be seen in Figure 3, shapes of the hypernym chains became more complicated as their length grows. Shorter chains (e.g. Length 3 and 4) formed a monotonically decreasing line, middle-length chains (e.g. 5, 6, and 7) showed a peak in the middle, and longer chains (e.g. 8 to 15) had more than one peak. Note that the four most common length chains shared a similar tendency in shape. They were length 5, 6, 7, and 8, and consisted 73% of the hyperny chains in the disambiguated set.

4.2 Parent-child pairs

The first part of our experiment has shown some general tendency of the relationship between document frequency and term specificity. We will now turn our focus to a smaller unit, the parent-child pair. More specifically, the second experiment aimed to investigate the number of cases where parent synsets held a higher document frequency than their children. It was also anticipated that this experiment should reveal the naïve performance of document frequency to identify a parent term for a given pair.

Table 2 shows the number of parent synsets holding a higher document frequency than child synsets at the next lower level. The number was measured based on the distinct pairs of parent and child synsets (in brackets) for each set of two levels.

There are two points that should be emphasised in the table. Firstly, the number of distinct pairs between two levels was higher at the more specific levels. This was due to the nature of the hierarchy where lower layers consisted of more words (or synsets) than the higher. Secondly, the highest probability of frequency determining specificity correctly (in **bold**) was almost always found in the lowest two levels. This indicated that document frequency was most reliable to

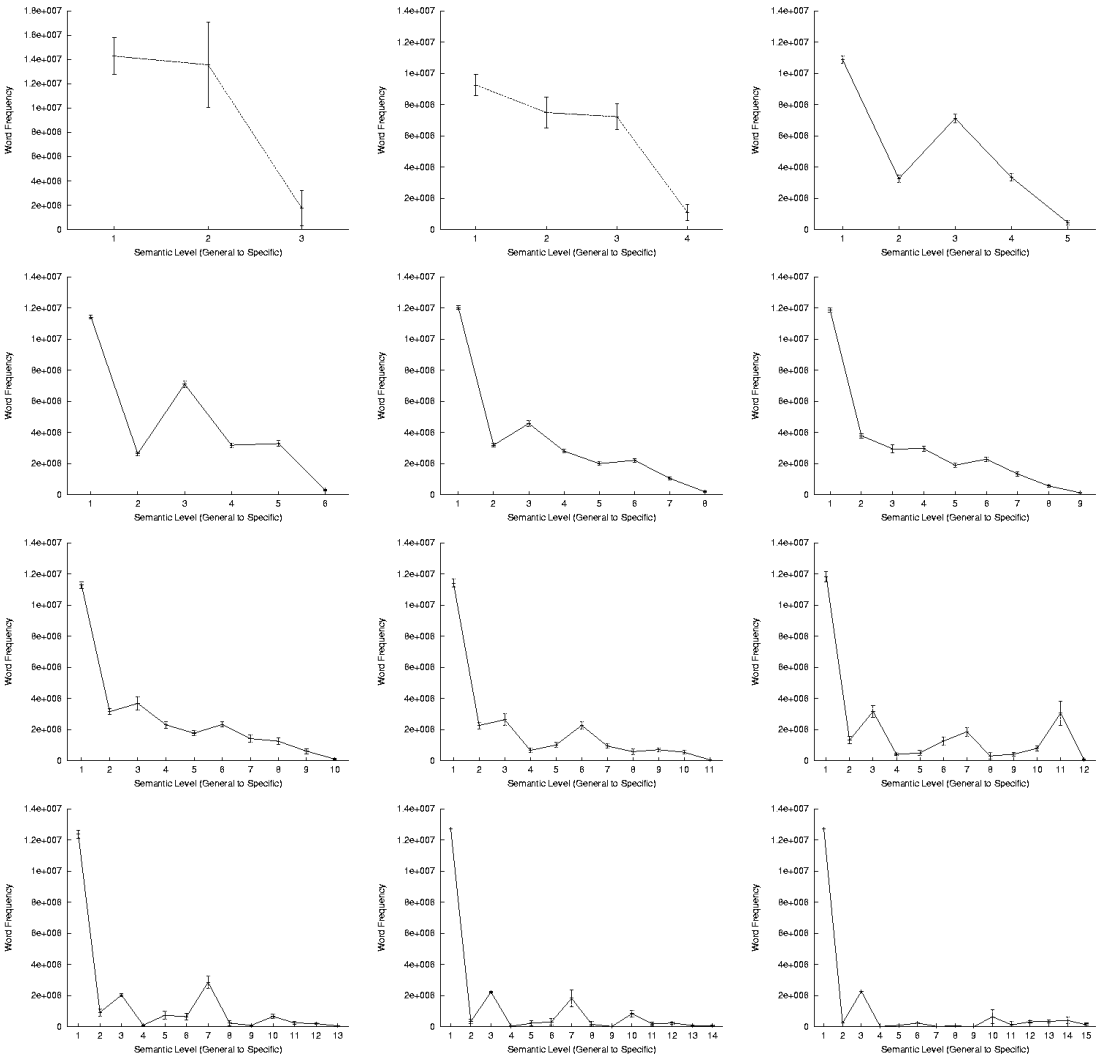


Figure 3: Average document frequency: Length 3 to 15

Table 2: Probability of parent synsets holding higher document frequency than the next lower child synsets

Chain length	No. of chain	Parent / Child	Semantic Level 1 / 2	2 / 3	3 / 4	4 / 5	5 / 6	6 / 7	7 / 8	8 / 9	9 / 10	10 / 11	11 / 12	12 / 13	13 / 14	14 / 15	Avg. (%)
3	157		17(28)	127(157)													
			60.71	**80.89**													70.80
4	733		15(27)	109(185)	547(673)												
			55.56	58.92	**81.82**												65.43
5	2696		13(23)	75(123)	316(486)	1789(2170)											
			56.52	60.98	65.02	**82.44**											66.24
6	5297		11(18)	45(79)	199(298)	592(910)	3252(4142)										
			61.11	56.96	66.78	65.05	**78.51**										65.68
7	5681		9(15)	29(50)	110(163)	253(417)	875(1269)	3705(4532)									
			60.00	58.00	67.48	60.67	68.95	**81.75**									66.14
8	4793		8(13)	18(30)	65(97)	116(201)	294(471)	810(1185)	3156(4033)								
			61.54	60.00	67.01	57.51	62.42	68.35	**78.25**								65.01
9	2773		6(10)	16(22)	33(53)	54(91)	105(190)	231(353)	488(808)	1957(2454)							
			60.00	72.73	62.26	59.34	55.26	65.44	60.40	**79.75**							64.40
10	1480		4(7)	11(16)	20(35)	28(50)	46(36)	69(90)	108(215)	273(486)	1040(1361)						
			57.14	68.75	57.14	56.00	46.75	69.23	50.23	56.17	**76.41**						59.76
11	863		3(5)	8(12)	13(18)	13(26)	16(37)	38(53)	44(90)	77(152)	167(300)	610(817)					
			60.00	66.67	72.22	50.00	43.24	71.70	48.89	50.66	55.67	**74.66**					59.37
12	368		3(4)	7(10)	9(12)	5(15)	7(19)	16(24)	14(37)	28(53)	37(72)	82(130)	235(336)				
			75.00	70.00	**75.00**	33.33	36.84	66.67	37.84	52.83	51.39	63.08	69.94				57.45
13	262		2(3)	2(3)	1(3)	0(4)	3(8)	7(10)	4(13)	9(16)	9(18)	18(28)	33(93)	186(262)			
			66.67	66.67	33.33	0.00	37.50	70.00	30.77	56.25	50.00	64.29	35.48	**70.99**			48.50
14	115		2(2)	1(2)	1(2)	0(2)	2(4)	3(4)	3(5)	4(6)	3(7)	8(11)	7(25)	31(48)	83(115)		
			100.00	50.00	50.00	0.00	50.00	75.00	60.00	66.67	42.86	72.73	28.00	64.58	72.17		56.31
15	24		1(1)	0(1)	1(1)	0(1)	0(1)	1(1)	1(2)	3(3)	0(3)	2(3)	2(4)	3(6)	6(10)	18(24)	
			100.00	0.00	**100.00**	0.00	0.00	**100.00**	50.00	**100.00**	0.00	66.67	50.00	50.00	60.00	75.00	53.69

identify more general terms from given pairs if both terms were very specific. This echoed the result of Caraballo and Charniak's work.

An overall performance of this task for 26,678 distinct pairs in WordNet was 75.78%. This accuracy is significant if one recalls that we only used document frequency data.

4.3 Effect of co-occurrence information

Up to this point, term specificity was determined through frequency of occurrence alone, however, it is often the case that the terms being measured co-occur with each other in a particular set of texts. Therefore, examining such collocated word pairs constituted the last part of our experiment: examining the impact of co-occurrence information when considering term specificity. Co-occurrence of terms was exploited in the context of selecting query expansion terms (Peat and Willet, 1991; Xu and Croft, 1996), or constructing a similarity thesaurus (Qiu and Frei, 1993). The information is generally considered to indicate the significance of contextual relatedness between terms. The more documents a given pair of terms share, the more the two terms are related.

To examine such effects, local document collections were built in the following manner. One hundred queries were taken from the Ad-Hoc task of TREC-6 (Voorhees and Harman, 1998) and TREC-7 (Voorhees and Harman, 1999) (Topics 301-400). For each query, a set of the top ranked 500 documents[7] was retrieved by an IR system [8] from the Financial Times (1991-94), LA Times (1989-89) and Wall Street Journal (1987-92) collections provided by TREC. As a result, a total of 300 sets (100 topics over the 3 newspapers) of the retrieved documents (referred to as *local collections*) were built and tested to observe the effect of co-occurrence information.

The test was carried out in the following way. Among the distinct 26,678 pairs in WordNet, those where both a parent and child term in synsets occurred in the 500 documents was recorded for

[7]Note that the reason why the size of a local collection was set to 500 was that the present experiments were part of a project that has been developing a multi-document summarisation system, and the system was designed to generate a summary from 500 documents.

[8]Developed by one of the authors

Table 3: Effect of co-occurrence information

	Without co-occs			With co-occs		
	Collection			Collection		
	FT	LA	WSJ	FT	LA	WSJ
No. of query	100	100	100	100	100	100
No. of pairs	122,913	166,712	155,421	9,951	15,824	14,246
Google DF	71.96%	72.52%	72.53%	83.37% (+15.86)	84.28% (+16.22)	83.35% (+14.91)
TREC DF	70.61%	71.28%	70.82%	83.21% (+17.84)	84.03% (+17.89)	83.43% (+18.71)
Top500 DF	65.13%	67.15%	66.11%	81.95% (+25.83)	83.15% (+23.83)	81.35% (+23.05)

each local set. Those pairs found to collocate in the documents were also recorded. For each type of recorded pair, the number of cases where parent terms held a higher document frequency than the child was counted. Document frequency was measured in three corpora: the 500 top documents, the newspaper collections from TREC (a super-set of the 500), and the Web pages indexed by Google. By using such corpora, it was possible to measure the impact of corpora size when determining specificity.

Table 3 shows the result of the test. There are several points to be made.

- The document frequency obtained from a larger size of corpus was found to be more accurate at determining specificity. This was unexpected. Our initial thought was that the local document frequency should provide a better performance since the local set of documents were likely to be more coherent than larger ones. However the result shows that document frequency from Google is best to identify more general (or specific) terms.

- The experimental results showed that it was easier to determine specificity from co-occurring term pairs. The impact of co-occurrence was found to be more significant when document frequency was obtained in a smaller size of documents. The improvement of accuracy was over 25% for the document frequency in the Top 500 documents but only 16% in document frequency determined from Google.

- The constancy of the result across the collections should be emphasised. As mentioned in Section 3.1, we were aware of the potential problem of WordNet's definition of commonest

Table 4: Coverage of vocabulary (Number of nouns found in collection)

WordNet	45,073
Google	45,055 (99.96%)
FT/LA/WSJ	23,705 (52.59%)
FT	18,366 (40.75%)
LA	20,225 (44.87%)
WSJ	18,497 (41.04%)

sense when considering the domain and heterogeneity of the tested corpora. If the senses of terms used in the Web and TREC Collections were significantly different, one would expect the results to have varied across the collections. However, our experiment indicated this was not the case. It would appear that on average, the usage of sense across the Web corpus was similar to sense usage in the newspaper corpora, which is perhaps surprising given the differences in age and domain of the corpora.

In addition, the coverage of vocabulary in these different collections was also analysed and shown in Table 4. As can be seen, the coverage of Google covers almost all of WordNet [9] while the TREC collections covers just over 50%. This result also highlights the usefulness of a large corpus such as the Web to determine term specificity.

5 Discussion and future work

This paper addressed several aspects of the relationship between document frequency and term specificity by using a significantly larger size of vocabulary and corpus than previous works. The series of experiments involved measuring average document frequency at various levels in hypernym chains, comparing of parent-child term pairs from WordNet, and evaluating the impact of co-occurrence information on determination of specificity.

[9] Examples of those terms appearing in WordNet's Noun Index but not in Google at the time of our experiment were *Hardenbergia comnptoniana* (a coral pea found in Western Australia), *Mimus polyglotktos* (a long-tailed songbird found in Southern U.S.), *Pisanosaur* (primitive dinosaur found in Argentina), and other professional academic terms or their alternative spelling.

From the first experiment, it was found that the most general and most specific level in hypernym chains were likely to hold the highest and lowest document frequency, respectively. However a monotonically decreasing line between them was not generally found. More often, one or more middle levels in the chains held the second (or third) highest document frequency in the chains. It is speculated that the terms in the middle level(s) belong to a basic level of category.

The second experiment focused on parent-child pairs in the chains to reveal the naïve performance of document frequency to identify a more general term from a given pair. The highest probability of parent synsets holding a higher document frequency than the next lower level of child synsets was found between the pairs of the last two levels in most cases. While a high probability was found between the first two levels, a wider range was found in the pairs between the middle levels. An overall performance of all distinct pairs defined in WordNet was 75.78%.

From these results, it is believed that document frequency can be used to determine term specificity more accurately when the terms are very specific. For the middle levels of specificity, document frequency still can be useful especially when the terms are more specific than the basic levels of category. The results also indicate that document frequency is less effective for the terms above the basic level.

The last experiment observed the effect of co-occurrence information on the probability of identifying a more general term of given pairs. The result showed 15% to 25% improvement in accuracy of the identification with the co-occurrence information. More improvement was added to document frequency obtained from local document set (i.e. Top 500 docs) than global set (i.e. Google). Although document frequency from a larger collection was found to be more accurate for the identification (due, it is assumed, to the bigger sample size of word occurrences), the result indicated the co-occurrence information can be useful where document frequency was only obtained from a small set of documents.

One of the challenges for future work will be to consider the impact of this study on term weight-

ing schemes used in IR. If as has been discussed, weighting schemes like IDF succeed because they are approximating term specificity, then consideration of the occasions when they fail in such an approximation may be beneficial. It may well be that IR systems are not well served by IDF calculated on words above the basic level in word hierarchies. Such an issue is likely to benefit from further examination.

Another direction in our research will be an investigation of the basic level of category and whether a method can be found to identify such a level when forming a concept hierarchy. More analysis of our data might lead us to produce a candidate list of terms belonging to the category.

Approaches to expand our study to achieve a better accuracy of determining specificity can be in many ways. One such approach can be illustrated as follows. Given that a co-occurrence based technique identifies the pairs of related terms, building a topic hierarchy based on a frequency based measure, then applying linguistically motivated techniques to those pairs that are likely to be less reliable (e.g. the pairs at higher levels in a hierarchy). For example, Hearst (1998) developed the simple pattern matchers to discover parent-child pairs using several key phrases: "X *such as* Y", "Y *and other* X", "X *including* Y", etc., where *X* can be seen as a parent of *Y*. The additional evidence from such a technique could be useful for improving the accuracy of specificity. If available, an online dictionary or encyclopedia can also be used. The advantage of using a frequency based measure to construct an initial hierarchy is that additional tools do not have to cope with a wide range of words but only those that word counting is likely to fail. Development of such an application is also one of our future work.

Acknowledgement

Thanks are due to Google who tolerated our many thousand queries, which allowed us to conduct the experiments. 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

Anick, Peter G. and Suresh Tipirneni. 1999. The Paraphrase Search Assistant: Terminological Feedback for Iterative Information Seeking. In M. Hearst, G. Gey, and R. Tong, editors, *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 153–161, Berkeley, CA.

Anick, Peter G. and Shivakumar Vaithyanathan. 1997. Exploiting Clustering and Phrases for Context-based Information Retrieval. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 314–323, Philadelphia, PA.

Barker, F. H., D. C. Veal, and B. K. Wyatt. 1972. Towards automatic profile construction. *Journal of Documentation*, 28(1):44–55.

Caraballo, Sharon A. and Eugene Charniak. 1999. Determining the specificity of nouns from text. In *Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 63–70.

Forsyth, R and R Rada. 1986. Adding an edge. In *Machine Learning: applications in expert systems and information retrieval*. Ellis Horwood, pages 198–212.

Grefenstette, Gregory. 1992. Use of syntactic context to produce term association lists for retrieval. In Nicholas J. Belkin, Peter Ingwersen, and Annelise Mark Pejtersen, editors, *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 89–97, Copenhagen, Denmark.

Grefenstette, Gregory. 1994. *Explorations in Automatic Thesaurus Discovery*. Kluwer.

Hearst, Marti A. 1998. Automated discovery of wordnet relations. In C Fellbaum, editor, *WordNet: An Electronic Lexical Database and Some of its Applications*. MIT Press, Cambridge, MA, pages 131–151.

Kilgarriff, Adam. 2001. Comparing corpora. *International Journal of Corpus Linguistics*. To appear, Available from `http://www.itri.brighton.ac.uk/~Adam.Kilgarriff/ijcl.pdf`.

Lakoff, George. 1987. *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*. University of Chicago Press, Chicago, IL.

Miller, George A. 1990. Nouns in wordnet: A lexical inheritance system. *International Journal of Lexicography*, 3(4):245–264.

Peat, Helen J. and Peter Willet. 1991. The limitations of term co-occurrence data for query expansion in document retrieval systems. *Journal of the American society for Information Science*, 42(5):378–383.

Qiu, Yonggang and H P. Frei. 1993. Concept based query expansion. In Robert Korfhage, Edie M. Rasmussen, and Peter Willett, editors, *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 160–169, Pittsburgh, PA.

Robertson, Steven E. 1974. Specificity and weighted retrieval. *Journal of Documentation*, 30(1):41–46.

Salton, Gerard. 1968. *Automatic Information Organization and Retrieval*. McGraw-Hill, New York.

Sanderson, Mark and Bruce Croft. 1999. Deriving concept hierarchies from text. In M. Hearst, G. Gey, and R. Tong, editors, *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 206–213, Berkeley, CA.

Sanderson, Mark and C. J. van Rijsbergen. 1999. The impact on retrieval effectiveness of skewed frequency distributions. *ACM Transactions on Information Systems*, 17(4):440–465.

Sparck Jones, Karen. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21.

Voorhees, E M. and D K. Harman, editors. 1999. *NIST Special Publication 500-242: the Seventh Text REtrieval Conference (TREC-7)*. NIST, Gaithersburg, MD.

Voorhees, Ellen and Donna Harman. 1998. Overview of the sixth text retrieval conference (TREC-6). In Ellen Voorhees and Donna Harman, editors, *NIST special publication 500-240: the sixth text retrieval conference (TREC 6)*, pages 1–24, Gaithersburg, MD. NIST.

Weinberg, Bella Hass and Julie A. Cunningham. 1985. The relationship between term specificity in mesh and online postings in medline. *Bulletin of the Medican Library Association*, 73(4):365–372.

Woods, William A. 1997. Conceptual Indexing: a Better Way to Organize Knowledge. Technical Report TR-97-61, Sun Microsoft Laboratories.

Xu, Jinxi and Bruce W. Croft. 1996. Query expansion using local and global document analysis. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 4–11, Zurich, Switzerland.

Appendix 5

Joho, H, Coverson, C, Sanderson, M. and Beaulieu, M. Hierarchical presentation of expansion terms. *In: Proceedings of the 17th ACM Symposium on Applied Computing, Madrid Spain, 2002*. New York: Association for Computing Machinery, 2002, 645-649.

Hierarchical Presentation of Expansion Terms

Hideo Joho, Claire Coverson, Mark Sanderson, and Micheline Beaulieu
Department of Information Studies, University of Sheffield
Western Bank, Sheffield, S10 2TN, UK
{h.joho, m.sanderson, m.beaulieu}@sheffield.ac.uk

ABSTRACT

Different presentations of candidate expansion terms have not been fully explored in interactive query expansion (IQE). Most existing systems that offer an IQE facility use a list form of presentation. This paper examines an hierarchical presentation of the expansion terms which are automatically generated from a set of retrieved documents, organised in a *general to specific* manner, and visualised by cascade menus. To evaluate the effectiveness of the presentation, a user test was carried out to compare the hierarchical form with the conventional list form. This shows that users of the hierarchy can complete the expansion task in less time and with fewer terms over those using the lists. Relations between initial query terms and selected expansion terms were also investigated.

Keywords

Information retrieval, interactive query expansion, concept hierarchies

1. INTRODUCTION

The increasing interest in providing online information available via the Internet has heightened the need for information retrieval (IR) systems that enable users to access heterogeneous resources that meet their information needs. The means of interaction between users and such an IR system to achieve a meaningful search is of particular interest and complexity. The potential benefit of interactive query expansion (IQE) has generated wide interest in making IR systems more adaptive as opposed to automatic query expansion (AQE), where users preferences are generally ignored.

An early form of interaction was through relevance feedback [18], where users judge the relevance of retrieved documents as indicative information about their interest given to the system. Harman [9] examined a more interactive approach by presenting a list of candidate expansion terms to users. The users then selected the terms of interest from the list to add to the initial queries. This interactive approach has been a standard way of presenting IQE and has been adopted by most researchers [1, 17, 7], regardless of the techniques used for extracting, or ranking the candidate terms.

SAC 2002, Madrid, Spain

The study of the user interface is also considered as an important aspect of IQE. Beaulieu [2] evaluated three different interface design based on the same IR system. In her study, a character-based design and two different graphical user interfaces were compared. This experiment highlights the importance of the close relationship between system functionality and interface design in supporting query expansion. Koenemann and Belkin [12] also showed that giving the user more control over how terms were added to the query could increase retrieval effectiveness and user satisfaction. A third study [13] compared the potential retrieved performance of IQE by simulating experienced and naive users.

As can be seen, researchers have addressed several aspects concerning the interaction between users and the IR systems in the context of IQE. However, little attention has been paid to the effectiveness of different presentation of candidate expansion terms. In fact, most experimental designs discussed so far merely use a form of list for the presentation.

The motivation of exploring a more effective presentation of expansion terms may resemble the studies of visualising search results. The existing IR systems including search engines on the Web present the search results as a list of document titles with some additional information such as URLs, size of documents, or a short summary. Users therefore have to spend some time scanning the list to locate relevant documents, and they are often reluctant to examine the documents ranked further below. Visualisation of the search results aims to help users to locate the documents of their interest regardless of ranking.

A number of techniques have been developed in this area [11, 24, 4], and Hearst [10] suggested giving users an overview as a means of navigation. She stated that 'an overview can help users get started, directing them into general neighbourhoods, after which they can navigate using more detailed descriptions' (p. 268). One of the types of overview is 'category hierarchies associated with the documents of a collection'. Subject directories exemplified by Yahoo![1] are such an example where topical concepts are organised in a hierarchical manner. Similarly, this type of navigation may be ideal for IQE.

In this paper, we examine the effectiveness of different presentation of expansion terms. The comparison is made between a hierarchical organisation and lists.

2. ORGANISATION AND VISUALISATION OF EXPANSION TERMS

In the previous section, it has been argued that existing IR systems have not paid much attention to *how to present expansion terms* [6], and the form of presentation is dominated by lists. Fur-

[1] `http://www.yahoo.com`

thermore, it is anticipated that a hierarchical presentation will provide users with an overview of candidate expansion terms, and thus, can be a promising alternative form.

Although manually constructed thesauri such as MeSH or INSPEC have been integrated into systems to assist in query formulation [20, 15], those resources are inevitably limited in the range of vocabulary and are not necessarily applicable to all domains.

A more promising technique that meets our aim has been introduced by Sanderson and Croft [19]. *Subsumption hierarchies* are designed to organise terms in a manner which is similar to existing manually constructed thesauri, or subject hierarchies in Yahoo.

Unlike other co-occurrence based techniques that measures *similarity* between terms [16], Sanderson and Croft use the co-occurrence information to identify a term that subsumes other terms. More specifically, a term, x, is said to subsume another term, y, if the documents which y occurs in are a subset of the documents in which x occurs. Given that a more frequent term tends to be more general [22], subsumption hierarchies organise terms in a 'general to specific' manner.

Although this technique was originally introduced as a means of automatic generation of concept hierarchies from a set of retrieved documents, applying it to IQE may also be valuable. For a comprehensive description of subsumption hierarchies, see [19].

The next section will describe a user test that examines the effectiveness of different presentation of expansion terms followed by the results and discussion.

3. USER TEST

The user test was carried out to examine any effects derived from different methods of presenting candidate expansion terms. In this study, a hierarchical presentation and conventional list presentation were compared.

3.1 Participants

A total of 24 subjects were recruited for the user test. The majority of the subjects (20) were students of the Department of Information Studies, University of Sheffield, and the rest were other members of the University. They consisted of 10 females and 14 males. The age of the subjects ranges from 22 to 35 with an average of 28.

3.2 Topics

Topics for the user test were taken from the TREC test collection (Topics 300-350) in the Sixth Text Retrieval Conference [23]. Among the topics, the subsumption hierarchy could not be created for topic 312, 330 and 348 as no relevance documents were retrieved in response to the queries, and therefore these topics were removed. It was also decided to remove topics which produced very small concept hierarchies. As a result, topics 316 and 327 which generated hierarchies of less than three levels or which contains less than 30 expansion terms were removed. A total of 45 topics remained and were used in the experiment.

3.3 Experimental system

INQUERY [3] was used as the IR system in this study. Candidate expansion terms were first extracted from the top 500 documents retrieved by INQUERY, in response to a query compiled from terms in the title of each topic description, then organised by the subsumption process, and finally visualised by the cascade menus. Lists were also generated using the identical set of terms included in the menus. The lists were ordered alphabetically as this was considered an arbitrary order for the presentation of the terms. An example of candidate expansion terms for topic 302 and terms selected by a test subject is presented in the Appendix.

3.4 Procedure

Subjects were first given an explanation of the reason behind the experiment. The subjects were told that a tool that attempts to generate a summary of the retrieved documents was in the process of being developed, and that query terms extracted from the documents retrieved would be shown on screen as an indication of what the retrieved document set is about.

The subjects were invited to consider the following scenario. They had just submitted a query to a retrieval system, the system had responded by showing a set of possible terms that could be added to the query in order to improve the search results. Their task would be to select terms they deemed appropriate to expand the query. Subject were then given a demonstration of the working system using a training query and topic description (topic 327) to illustrate the procedure.

After the training session, they were asked to carry out the actual experimental expansion task with nine topics. In order to save time, all menus and lists for the 45 topics were generated in advance of the user test. Subjects were alternately assigned to a control or experimental group. The experimental group was presented with the interface containing the menus (the Menu group) and the control group was presented with the lists (the List group).

Following the completion of the expansion tasks by all the participants, selected expansion terms were added to the initial query for each topic, and the search was re-run with the expanded queries. A comprehensive description of the user test can be found in [5].

4. RESULTS AND DISCUSSION

This section describes experimental results of the user test based on the standard precision and recall measures, the number of expansion terms selected, the time taken to complete the task, and link type analysis of expansion terms.

4.1 Precision and Recall

A recall-precision graph (Figure 1) was plotted using the top 1000 retrieved documents retrieved in response to the unexpanded queries, expanded queries by lists, and by menus. The measurements of recall and precision were based on the TREC relevance assessments [23]. The graph shows the unexpanded queries retrieve documents at a higher precision for lower recall while the

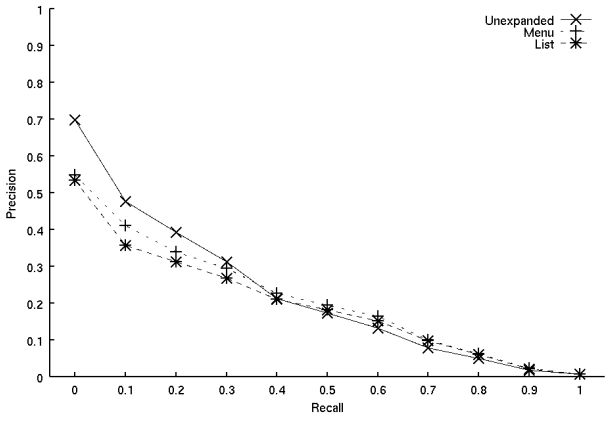


Figure 1: Precision-Recall graph

expanded queries produce a higher precision for higher recall after the 0.4 point. The graph also shows that the queries expanded by the menus constantly perform slightly better than the lists. No statistical significance was, however, found between them.

Table 1 shows precision at the top one, five, ten and twenty retrieved documents. As indicated in the precision-recall graph, the unexpanded queries retrieve more relevant document at all the levels shown than the expanded queries. Similarly, precision of the expanded queries by the Menus are higher than the List except at level one.

Table 1: Precision at one, five, ten and twenty

		Expanded		Residual	
	Unexp.	List	Menu	List	Menu
PREC-AT-1	0.6117	0.4078	0.4078	0.2816	0.2330
PREC-AT-5	0.4252	0.3612	0.3767	0.2447	0.2485
PREC-AT-10	0.3903	0.3097	0.3408	0.2223	0.2223
PREC-AT-20	0.3393	0.2937	0.2966	0.2155	0.2044

The residual precision (Fifth and Sixth columns in Table 1) is calculated by a set of documents retrieved by the expanded queries but removing the relevant documents that the corresponding initial query already retrieved in the top 20. The lower ranked documents were then promoted. This simulates the performance of the expanded queries where a user is supposed to identify all the relevant documents from initial search results in the top 20. This also reveals the extent to which new relevant documents were retrieved by the expanded queries.

The data shows that both expanded queries cause new relevant documents to be retrieved in the top 20. However, there is little difference between the lists and menus in terms of the standard retrieval effectiveness shown above.

4.2 Number of expansion terms and time to complete task

The standard retrieval effectiveness was shown in the previous section. However, the standard method may not be fully adequate for evaluating the line of research presented here [21]. This section presents data concerning efficiency to complete the expansion task. Number of expansion terms selected by subjects and time to complete the whole expansion task are shown in Table 2. They can indicate the effectiveness of term selection tasks frequently carried out in IQE.

Table 2: Efficiency of expansion task

	List	Menu
No. of selected expansion terms	16.87	12.83
Time to complete expansion task (sec.)	203.28	168.40

As can be seen, the subjects in the Menu group completed the task with over four terms fewer than the List group on average, and this significantly shortened the time to complete the task . These two performance figures are statistically significant ($p < 0.001$).

4.3 User perceptions

In order to gain an indication of the ability of the interfaces to provide an overview of the retrieved documents, subjects in both groups were asked if, after exploring the expansion terms, they felt that they had a better idea of the contents of the retrieved documents. 80% of the subjects in the Menu group stated that they had a better idea of the contents. This is significantly higher than the List group where only 42% felt that exploring the list gave them a better idea of the contents. One subject in the Menu group also argued that the hierarchies gave them an idea of whether or not they were going to find a decent answer to their query. This seem to support our approach to provide an adequate summarisation of documents retrieved in response to a query and that the subsumption hierarchies are meaningful.

4.4 Link types

The semantic and/or topical relations between initial query terms and expansion terms have been of interest in IQE since this indicates an aspect of user's searching behaviour [7, 8].

The relations used in our link type test were based on the ones defined by WordNet [14], but an additional relation, *conceptually related* (or contextually related), was also devised. A conceptually related term was defined more broadly than the related term (RT) in an existing thesaurus. An example of such a relation is *tooth* and *dentist*. This type of relation is not defined in WordNet but can be useful in the context of query expansion [20].

Table 3: Link types between initial query terms and expansion terms selected by the Menu group (N=131)

Relation	Portion (%)
Hyponym	8
Meronym	2
Hypernym	9
Holonym	0
Coordinate Sister	2
Synonym	13
Antonym	0
Conceptually Related	65
Other/Don't know	1
Total	100

Table 3 shows the link types between initial query terms and expansion terms selected by the Menu group. This reveals that far more than half of the expansion terms selected by the subjects were terms conceptually related to the initial query terms. Synonyms were the second largest portion among the relations. A total of 10% of the selected expansion terms were of aspects of corresponding initial query terms (Hyponym and Meronym). A similar portion was also found in the parental relation (Hypernym) with the initial query terms.

Arguably, this result can be seen as an echo of the findings from Efthimiadis's experiment [7] where 44% of selected expansion terms were not considered as a type of thesaurus-like relations but as *new ideas* by the users. A similar tendency was found in our test. The reason for this tendency is not clear [7], but it is speculated that many terms which users choose as further descriptions of their information needs can be more complex (or flexible) relations to initial queries than those in a thesaurus.

This also highlights the advantage of the subsumption hierarchies as a means of hierarchical presentation of candidate expansion terms, as opposed to a manually constructed thesaurus which provides well organised but limited range of relations.

5. CONCLUSION

This paper examined the effectiveness of the hierarchical presentation of candidate expansion terms in comparison of a conventional list form. Although no significant difference in terms of

precision-recall between them was found, the users in the hierarchical presentation group completed the expansion task in significantly shorter time than users in the list presentation group. From these results, it is concluded that different presentations of expansion terms do effect the expansion task, and a more structured presentation can improve the selection of expansion terms. This also suggests that more attention on presenting expansion terms should be made in the research of IQE.

6. ACKNOWLEDGMENTS

Tim Gollins helped in postprocessing the experimental data. This work was carried out as a part of the CiQuest project, which is 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.

7. REFERENCES

[1] J. Allan, L. Ballesteros, J. Callan, W. Croft, and Z. Lu. Recent experiments with InQuery. In D. K. Harman, editor, *Proceedings of the Fourth Text REtrieval Conference (TREC-4)*, pages 49–72, Gaithersburg, MD, 1996. NIST.

[2] M. Beaulieu. Experiments on interfaces to support query expansion. *Journal of Documentation*, 53(1):8–19, 1997.

[3] J. P. Callan, B. W. Croft, and S. M. Harding. The INQUERY retrieval system. In *Proceedings of the Third International Conference on Database and Expert Systems Applications*, pages 78–83, Valencia, Spain, 1992. Springer-Verlag.

[4] H. Chen and S. Dumais. Bringing order to the Web: Automatically categorizing search results. In *Proceedings of the CHI 2000 Conference on Human factors in computing systems*, pages 145–152, The Hague Netherlands, 2000. ACM.

[5] C. Coverson. *Query Expansion Using an Interactive Concept Hierarchy*. Master's dissertation, Department of Information Studies, University of Sheffield, 2000.

[6] E. N. Efthimiadis. Query expansion. *Annual Review of Information Systems and Technology*, 31:121–187, 1996.

[7] E. N. Efthimiadis. Interactive query expansion: A user-based evaluation in a relevance feedback environment. *Journal of the American Society for Information Science*, 51(11):989–1003, 2000.

[8] J. Greenberg. Optimal query expansion (QE) processing methods with semantically encoded structured thesauri terminology. *Journal of the American Society for Information Science and Technology*, 52(6):487–498, 2001.

[9] D. Harman. Towards interactive query expansion. In *Proceedings of the 11th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 321–331, Grenoble, France, 1988. ACM.

[10] M. A. Hearst. User interfaces and visualization. In R. Baeza-Yates and B. Ribeiro-Neto, editors, *Modern Information Retrieval*, pages 257–323. ACM Press, New York, 1999.

[11] M. A. Hearst and J. O. Pederson. Re-examining the cluster hypothesis: Scatter/Gather on retrieval results. In H.-P. Frei, D. Harman, P. Schable, and R. Wilkinson, editors, *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 76–84, Zurich, Switzerland, 1996. ACM.

[12] J. Koenemann and N. J. Belkin. A case for interaction: A study of interactive information retrieval behavior and effectiveness. In *Conference Proceedings on Human Factors in Computing Systems (CHI '96)*, pages 205–212, Vancouver, Canada, 1996. ACM.

[13] M. Magennis and C. J. van Rijsbergen. The potential and actual effectiveness of interactive query expansion. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 324–332, Philadelphia, PA, 1997. ACM.

[14] G. A. Miller. WordNet: A lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995.

[15] S. Pollitt. Interactive information retrieval based on faceted classification using views. In *Proceedings of the 6th International Study Conference on Classification (FID/CR)*, London, UK, 1997. University College of London. Available from `http://www.hud.ac.uk/schools/cedar/dorking.htm` [Accessed: 29/03/2001].

[16] Y. Qiu and H. P. Frei. Concept based query expansion. In R. Korfhage, E. M. Rasmussen, and P. Willett, editors, *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 160–169, Pittsburgh, PA, 1993. ACM.

[17] S. E. Robertson, S. Walker, and M. Beaulieu. Laboratory experiments with Okapi: Participation in the TREC programme. *Journal of Documentation*, 53(1):20–34, 1997.

[18] G. Salton and C. Buckley. Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, 41(4):288–297, 1990.

[19] M. Sanderson and B. Croft. Deriving concept hierarchies from text. In M. Hearst, G. Gey, and R. Tong, editors, *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 206–213, Berkeley, CA, 1999. ACM.

[20] B. R. Schatz, E. H. Johnson, P. A. Cochrane, and H. Chen. Interactive term suggestion for users of digital libraries: Using subject thesauri and co-occurrence lists for information retrieval. In *Proceedings of the 1st ACM International Conference on Digital Libraries*, pages 126–133, Bethesda, MD, 1996. ACM.

[21] M. M. Sebrechts, J. Vasilakis, M. S. Miller, J. V. Cugini, and S. J. Laskowski. Visualization of search results: A comparative evaluation of text, 2D, and 3D interfaces. In M. Hearst, G. Gey, and R. Tong, editors, *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3–10, Berkeley, CA, 1999. ACM.

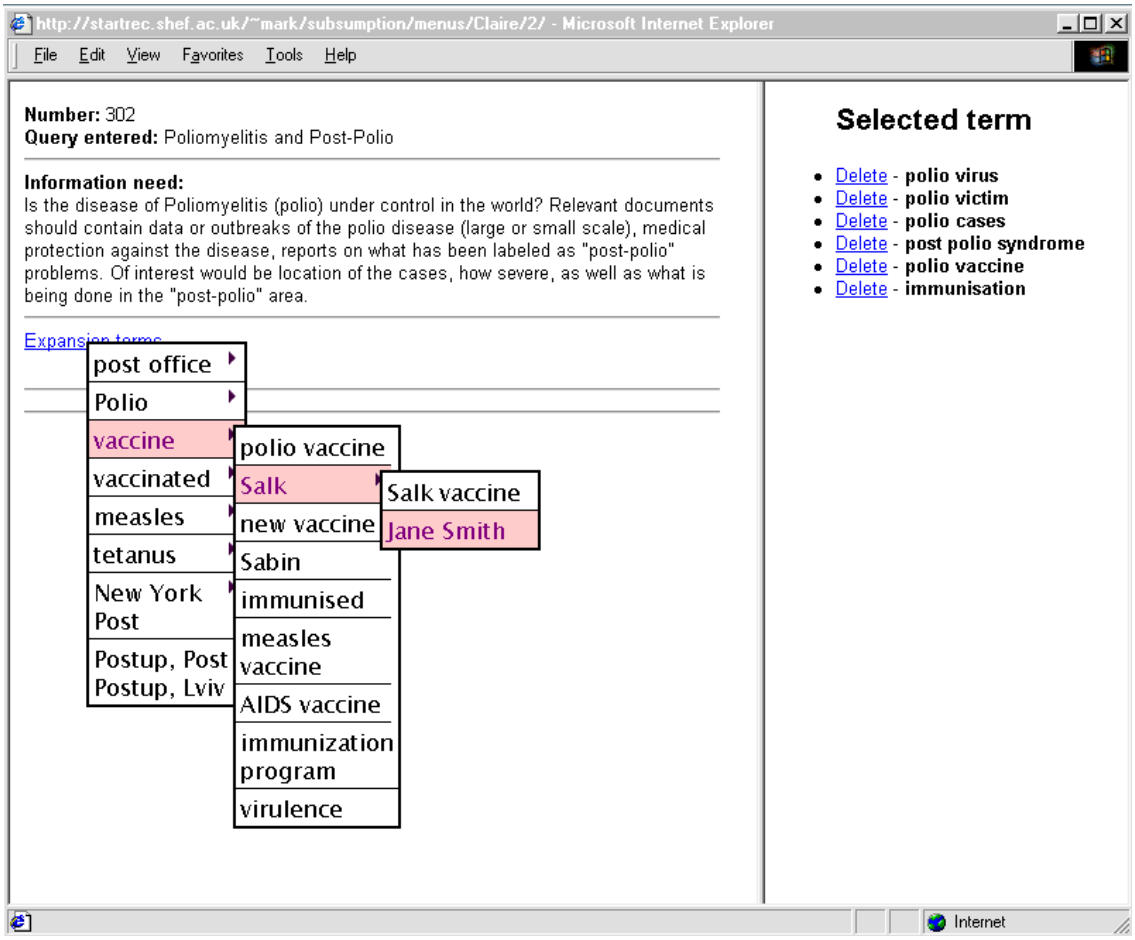
[22] K. Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21, 1972.

[23] E. Voorhees and D. Harman. Overview of the sixth text retrieval conference (TREC-6). In E. Voorhees and D. Harman, editors, *NIST Special Publication 500-240: The Sixth Text REtrieval Conference (TREC 6)*, pages 1–24, Gaithersburg, MD, 1998. NIST.

[24] O. Zamir and O. Etzioni. Grouper: A dynamic clustering interface to web search results. In A. Mendelzon, editor, *Proceedings of the Eighth International World Wide Web Conference (WWW8)*, Toront, Canada, 1999. Elsevier Science.

APPENDIX

A. SAMPLE USER INTERFACE



Appendix 6

Joho, H, Sanderson, M and Beaulieu, M. Hierarchical approach to term suggestion device. In*: Proceedings of the 25th Annual ACM SIGIR Conference on Research and Development in Information Retrieval, Tampere, Finland. 2002*. New York: Association for Computing Machinery, 2002, p.454.

Hierarchical Approach to Term Suggestion Device

Hideo Joho
Department of Information Studies
University of Sheffield
Western Bank, Sheffield, S10 2TN
+44 (0)114 222 2675
h.joho@sheffield.ac.uk

Mark Sanderson
Department of Information Studies
University of Sheffield
Western Bank, Sheffield, S10 2TN
+44 (0)114 222 2648
m.sanderson@sheffield.ac.uk

Micheline Beaulieu
Department of Information Studies
University of Sheffield
Western Bank, Sheffield, S10 2TN
+44 (0)114 222 2640
m.beaulieu@sheffield.ac.uk

ABSTRACT

Our demonstration shows the hierarchy system working on a locally run search engine. Hierarchies are dynamically generated from the retrieved documents, and visualised on the menus. When a user selects a term from the hierarchy, the documents linked to the term are listed, and the term is then added to the initial query to rerun a search. Through the demonstration we illustrate how hierarchical presentation of expansion terms is achieved, and how our approach supports users to articulate their information needs using the hierarchy.

Keywords

Concept hierarchies, interactive query expansion.

1. DESCRIPTION

Traditionally term suggestion devices in Interactive Query Expansion (IQE) have been designed to present candidate expansion terms in a form of list, regardless the underlying ranking techniques. While such a list has been the dominant form of presentation, it seems less effective to provide users with the contexts from which the terms are derived.

To explore an alternative approach to support IQE, we have been investigating methods of presenting expansion terms in a hierarchical form: automatically and dynamically derived from a set of retrieved documents; terms (concepts) are organised based on term specificity, with allowing links to contextually related terms (e.g. Yahoo Directory). It is anticipated that such a hierarchy will provide users with an overview of documents, and can be useful to allow users to articulate their information need.

In our demonstration, we use a technique called "subsumption hierarchy" [1] to construct a concept hierarchy. The subsumption hierarchy uses the co-occurrence information to identify a pair of terms that are related, and measures term specificity using document frequency. Although this is a simple technique, it is encouraging that the technique takes account into term specificity in constructing hierarchies. A user test has shown a positive reaction of the hierarchy in the expansion task [2].

Our live demonstration shows the hierarchy system working on a locally run search engine. Hierarchies are dynamically generated from the retrieved documents, and visualised on the menus (see Figure 1: the number in brackets beside terms indicates the number of documents linked to the terms). When a user selects a term from the hierarchy, the documents linked to the term are listed, and the term is then added to the initial query to rerun a search. Through the demonstration we illustrate how hierarchical presentation of expansion terms is achieved, and how our approach will support users to articulate their information needs using the hierarchy. We also briefly discuss possible ways to improve the current system.

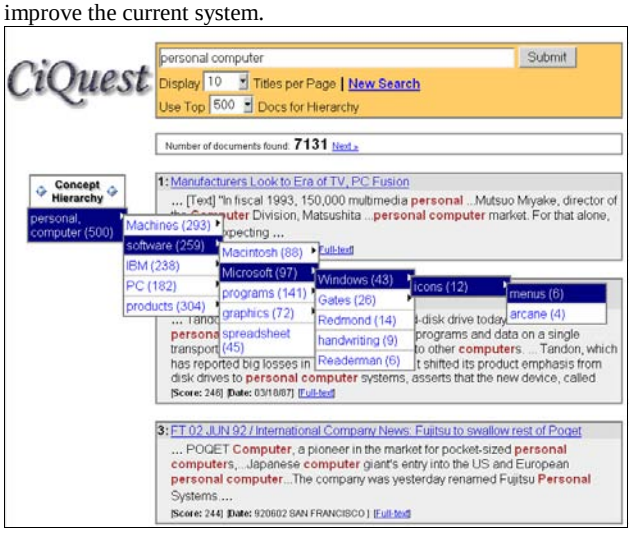


Figure 1. Screenshot of CiQuest System: a sample hierarchy generated from the top 500 documents retrieved in response to a query "personal computer". The maximum number of terms at each level was truncated to five for this screenshot.

2. ACKNOWLEGEMENT

This work was carried out as a part of the CiQuest project, which is 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.

3. REFERENCES

[1] Sanderson, M. and Croft, B. Deriving Concept Hierarchies from Text. In: *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 206-213, Berkeley, CA, 1999, ACM.

[2] Joho, H., Coverson, C., Sanderson, M. and Beaulieu, M. Hierarchical Presentation of Expansion Terms. In: *Proceedings of the 17th ACM Symposium on Applied Computing (SAC'2002)*, 645-649, Madrid, Spain, 2002, ACM.

SIGIR '02, August 11-15, 2002, Tampere, Finland.
ACM 1-58113-561-0/02/0008.

Appendix 7

Joho, H, Sanderson, M and Beaulieu, M. A study of user interaction with a concept-based interactive query expansion support tool. In*: Proceedings of the 26th European Conference on Information Retrieval,* Sanderland, UK. 2004. Lecture Notes in Computer Science, Berlin: Springer-Verlag, 2004.

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.

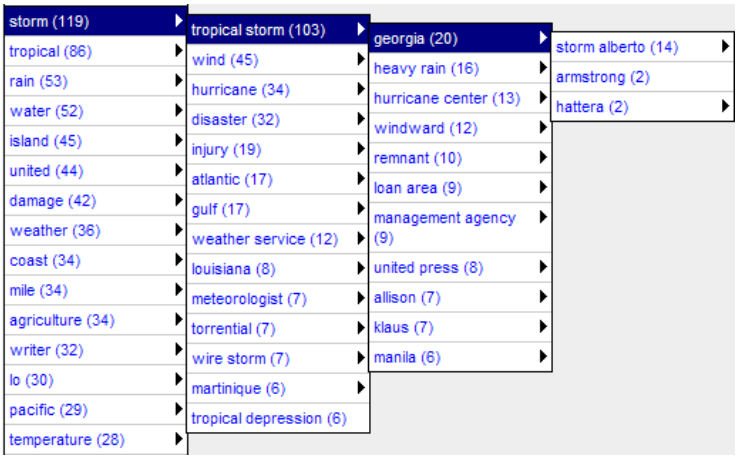


Fig. 2. Sample hierarchy generated by the subsumption approach with the top 200 documents retrieved in response to the query *tropical storm*. The number next to the term indicates the frequency of occurrence. You can see the phrase ”tropical storm” is subsumed by the term ”storm”. Also several instances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

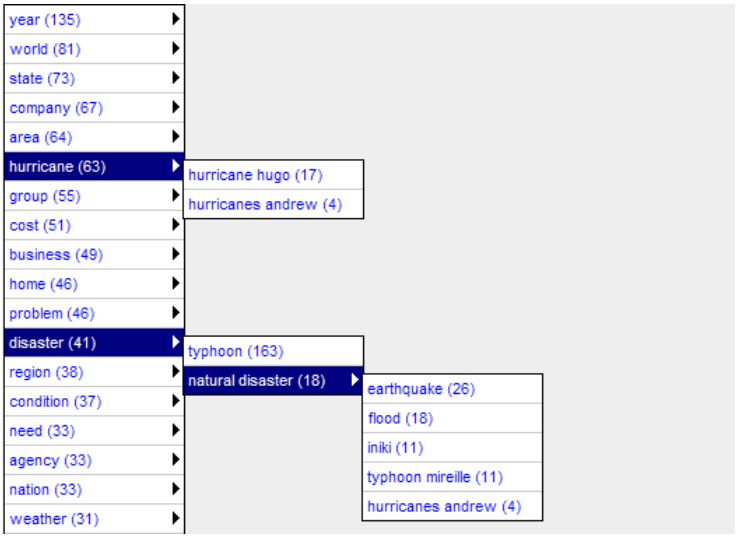


Fig. 3. Sample hierarchy generated by the trigger phrase approach with the top 200 documents retrieved in response to the query *typhoon hurricane*. Noun phrases such as *hurricane hugo* and *hurricane andrew* can be found under the head noun ”hurricane” at the top level of the hierarchy. Also you can 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.

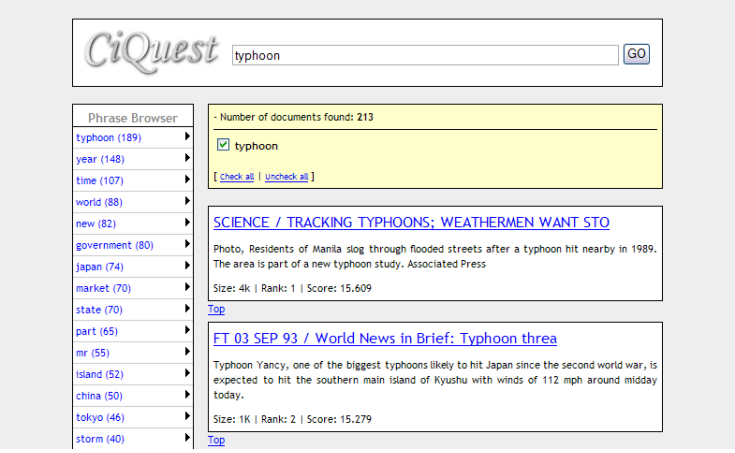


Fig. 4. CiQuest system: Top level of menu is shown along with the search result

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	
Topic ID	Official instances	Instance recall	Instance precision	Instance recall	Instance precision	Instance recall	Instance 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