# Supporting user query reformulation and searching: A concept hierarchy approach

A study submitted in partial fulfilment

of the requirements for the degree of

Doctor of Philosophy

At


The University Of Sheffield.

By

**Hideo Joho**

*January 2007*

# Abstract

While technological advances have enabled us to access extensive document collections, formulating a query which is well designed for an information retrieval (IR) system remains a difficult task. A number of methods have been developed to support user query formulation and reformulation based on terminological feedback. Terminological feedback offers a set of terms that can be used to modify an existing query. This also gives users an opportunity to transform a part of the query re/formulation process to term selection: a potentially simpler task. There is, however, much room for investigating and improving the interaction between users and IR systems with regard to query re/formulation. The limited context and structure of suggested terms are just some of the problems found with existing methods.

This thesis presents a new approach to supporting user query reformulation and searching. The approach is based on hierarchical organisation of terms, which is dynamically derived from a set of retrieved documents. This thesis investigates both statistical and lexical aspects of terms as a means of deriving a hierarchy from texts. As a summative evaluation of our approach, a user study is carried out to investigate several aspects of the user interaction with our support system. A search interface is developed to integrate a visualised hierarchy into a search result of an IR system. Two types of hierarchies are evaluated based on a TREC test collection, and compared to a baseline that has no hierarchy. Results suggest that multiple aspects of information searching process can be supported by the hierarchies. In particular, the range of search vocabulary employed to complete a task is shown to increase, and browsing of retrieved documents is found to be facilitated by the hierarchies.

# Preface

This thesis contains a revised version of the materials that have been published elsewhere in the following publications.

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*, McLean, VA. pp. 180-186. 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*, San Diego, CA. pp. 219-221. Morgan Kaufmann.

Joho, H., Coverson, C., Sanderson, M., and Beaulieu, M. (2002). "Hierarchical Presentation of Expansion Terms". In: Lamont, G.B., Haddad, H., Papadopoulos, G., and Panda, B. (eds.), *Proceedings of the 17th ACM Symposium on Applied Computing*, Madrid, Spain. pp. 645-649. ACM.

Joho, H., Sanderson, M., and Beaulieu, M. (2002). "Hierarchical Approach to Query Suggestion Device". In: Beaulieu, M., Baeza-Yates, R., and Myaeng, S.H. (eds.), *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Tempere, Finland. pp. 454. ACM.

Joho, H., Sanderson, M., and Beaulieu, M. (2004). "A Study of User Interaction with a Concept-based Interactive Query Expansion Support Tool". In: McDonald, S. and Tait, J. (eds.), *Advances in Information Retrieval, 26th European Conference on Information Retrieval*, Vol. 2997, pp. 42-56. Sunderland, UK: Springer.

# Acknowledgements

I would like to thank the following people for their support during my PhD.

First of all, I thank my supervisors, Dr. Sanderson and Prof. Beaulieu, for their patient support, encouragement, and friendly relationship throughout the project. I am truly fortunate to have you two as my supervisors. My examiners, Prof. Harper (Robert Gordon University) and Prof. Whittaker (University of Sheffield), helped me improve this thesis.

I thank participants of my experiments in the Department of Information Studies and Computer Science for their time and feedback.

I also thank the current and past members of the IR group at the Department of Information Studies, for having stimulating discussions on various topics and for reading the drafts of this thesis. Anna, Asaad, Daniela, Helene, Kelly, Paul, Simon, Stephen, and Xiao Mang, thank you.

My thanks also go to the current and past members of the Department of Information Studies: Mariano, thank you for your wonderful Spanish dishes; Miguel and Jośe, thank you for your wonderful Porto wines; Andrew, thank you for your introduction to wonderful British (and Irish) beers.

I also thank Prof. Matsumura (Professor Emeritus, University of Library and Information Science, Japan), Prof. Tamura (Keio University, Japan), and Prof. Naito (Toyo University, Japan) for their support in my application to the University of Sheffield.

My thanks also go to Dr. Jose and Prof. Van Rijsbergen (University of Glasgow) for kindly allowing me to have some time off the project to complete my PhD.

Finally, I thank my parents for giving me an opportunity to study abroad.

This work was supported by the CiQuest project funded by the Library and Information Commission (now the Museums, Libraries and Archives Council), the SPIRIT project funded under EC Fifth Framework Program (Contract ref: IST-2001-35047), and the Adapt project funded by EPSRC (Contract ref: EP/C004108/1). Any opinions, findings, and conclusions described here are the author's and do not necessarily reflect those of the sponsors.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

## 1.1 Research problem

Formulating a query which adequately represents an underlying information need has been known as one of the most difficult tasks for a user of an Information Retrieval (IR) system. While the limited knowledge of a topic can be a motivation of accessing an IR system, it can cause a user to formulate a short query consisting of broad terms (Jansen et al., 1998; Silverstein et al., 1998). The short queries, including well represented ones, are likely to achieve a worse retrieval performance than longer queries across various topics (Voorhees and Harman, 1998). Although one way to improve the performance is to reformulate the initial query, it has been shown that people have rarely modified their queries manually (Spink et al., 2001).

Query expansion (QE) is a technique to supplement an initial query with additional terms, and has long been studied in IR as a means of improving retrieval effectiveness (Efthimiadis, 1996). One of the most popular strategies of QE is based on relevance feedback (Salton and Buckley, 1990). Relevance feedback (RF) refers to a process of obtaining relevance information about retrieved documents. It typically either asks a user to indicate a set of documents s/he thinks are relevant (Robertson et al., 1995a), or assumes that the top N ranked documents are relevant (Salton and Buckley, 1990). An initial query can be expanded by adding the terms found in perceived relevant documents.

An alternative approach to QE is based on term clustering, which is a statistical

technique that groups together similar terms in a multidimensional space (Sparck Jones, 1971). Term clustering is often motivated by the association hypothesis, which states "if an index term is good at discriminating relevant from non-relevant documents, then any closely associated index terms is also likely to be good at this" (p.134) (Van Rijsbergen, 1979). The similarity between terms is, for example, estimated based on the number of documents in which pairs of terms co-occur in the entire document collection (Salton, 1980). An initial query can be expanded by adding the member terms in a cluster to which query terms belong.

Experimental results suggest, however, that the success of these automatic QE techniques often depends on the quality of initial queries and their search results. While a large improvement can be achieved when a sufficient number of relevant documents are retrieved by an initial query, the performance is degraded when limited relevant documents are available for expansion (Magennis and van Rijsbergen, 1997). This somehow contradicts to the potential benefit of query expansion since a user is likely to appreciate an automatic improvement when the search result is not satisfactory.

More recently, there has been growing interest in the possibility of reformulating queries interactively. Unlike the automatic QE, the main aim is to assist users in reformulating existing queries by suggesting a set of candidate terms (Harman, 1988). From a user's point of view, this is a form of terminological feedback from an IR system, and it could transform a part of query reformulation process to term selection: a potentially simpler task. Therefore, the area of interactive query reformulation not only pursues the improvement of retrieval performance through the development of a better term suggestion method, but also it involves the investigation of information seeking behaviour, human-computer interaction, and user interface design with regard to query reformulation process (Beaulieu, 2000).

While terminological feedback has been implemented and integrated into early IR systems (Doszkocs, 1983; Ingwersen, 1992; Wade and Willett, 1988), few of them have evaluated the effectiveness in terms of a support of user query reformulation. A large scale user test was carried out by Koenemann and Belkin (1996), and an interactive QE (IQE) approach based on RF was evaluated. The experimental results show that a greater level of control given to a user in query expansion can lead to

a better retrieval performance as well as user satisfaction. This work has provided a promising evidence which suggests that IQE can be effective for supporting user query reformulation.

However, other studies have indicated that a mere conversion of automatic QE techniques into an interactive fashion does not necessarily help users make an adequate selection of useful terms (Beaulieu, 1997; Belkin et al., 1999; Magennis and van Rijsbergen, 1997; Ruthven, 2003). A typical approach to IQE is to offer a small set of candidate terms which are ranked by an underlying automatic QE method. However, statistically estimated and ranked terms can be too specific or unfamiliar for a user to recognise the terms' usefulness (Belkin et al., 1999). Therefore, a user appears to select some terms only because they are recognisable (Ruthven, 2003). Moreover, there could be a large number of terms that are potentially relevant to an underlying information need (Anick and Tipirneni, 1999); a short list of ranked terms is not always ideal to present a wide range of related terms.

To address these problems, an alternative approach, which is different from existing ranking-based techniques, should be devised to support user query reformulation. Such an approach should ideally offer a range of related terms presented in a structured way, so that a system's terminological feedback provides sufficient contextual information for a user to modify their queries effectively.

## 1.2 Research aim and objectives

This thesis aims to address some of these issues by introducing a concept-based approach to query reformulation and searching. This thesis investigates and evaluates automatic methods of deriving a hierarchical structure of concepts from retrieved documents, to support a user's query reformulation and searching. The main objective is 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 specific objectives includes:

- investigating methods for the automatic generation and hierarchical organisation of concepts derived from retrieved documents;

- exploring the presentation of document derived concept structures for incorporation into a user interface;
- evaluating the effectiveness of statistical and lexical aspects of concepts to determine hierarchical relationships;
- evaluating the retrieval effectiveness of document derived concept structures for selecting relevant documents in a retrieved document set;
- evaluating the retrieval effectiveness of incorporating concept structures to assist users in selecting candidate terms for query reformulation; and finally,
- investigating how searchers make use of concept structures to perform a search task with an interactive IR system.

The overall methodological approach taken in this thesis is to address the interdependent research issues related to the stated objectives in a progressive and integrated fashion. A special emphasis is put on user participation in the investigation where the approach takes full account of user searching behaviour and the interactive searching process as well as retrieval effectiveness in the evaluation.

## 1.3 Thesis outline

This thesis is organised as follows.

Chapter 2 introduces the basic concepts of information retrieval that are related to this thesis. The indexing, ranking models, and systematic evaluation measurements are discussed.

Chapter 3 reviews studies related to query reformulation. The motivation of query reformulation is discussed first followed by the investigation of existing techniques of query expansion. Issues of user interaction with candidate terms are also discussed. The limitations of existing approaches are addressed and a new approach is proposed.

Chapter 4 reviews studies related to concept hierarchies which form the basis of our approach. Both statistical and lexical techniques to derive a hierarchical rela-

tionship of concepts are investigated. The presentation of concept structures is also discussed.

Chapter 5 provides an overview of experiments presented in this thesis. This chapter discusses how IR test collections are used in our corpus-based studies and a task-based user study.

Chapter 6 presents the first experiment which aims to study the relationship between the document frequency and term specificity. A series of large scale experiment is carried out using a thesaurus and varied sizes of document collections. The chapter aims to provide further insights into document frequency's relatedness to term specificity. The effect of co-occurrence information is also examined.

Chapter 7 presents the second experiment which aims to evaluate pattern matching techniques to extract parent-child descriptions from texts. Lexicon-syntactic patterns are expanded and combined with other corpus-based statistical evidences to rank candidate descriptive phrases of concepts.

Chapter 8 presents a user study as the summative evaluation of our approach to supporting a user's query reformulation and searching. A search interface is developed where a visualised hierarchy is integrated into a search result of an IR system. The main aim of the chapter is to investigate the user interaction with our interface and to measure the retrieval effectiveness in a task-based user study.

Finally, Chapter 9 concludes the thesis by highlighting the major findings of the present work. The potential directions of future work are also discussed.

# Chapter 2

# Basic concepts of Information Retrieval

## 2.1 Introduction

The field of information retrieval (IR) is primarily concerned with storage and retrieval of unstructured objects (Salton and McGill, 1983). Therefore, an IR system can be seen as a system that stores a collection of unstructured objects and retrieves a part of the collection in response to a request. While many IR systems are designed to process textual data (Voorhees and Buckland, 2004), other types of media such as images and movies have also been investigated (Enser et al., 2004). The aspect of cross-language (i.e. retrieving an object represented by a different language from a query) is equally an active area in IR (Braschler and Peters, 2004). Technologies have also been applied to semi-structured objects such as XML documents (Fuhr et al., 2004). This chapter introduces some of the basic concepts of IR which are relevant to this thesis.

Let us begin with an abstract picture of the environment where an IR system might be used. Figure 2.1 illustrates a simplified view of information access process using an IR system. From the left, there is a user who has a potentially intangible information need. The user formulates a query which somehow represents the information need. The query is submitted to an IR system which has an access to a collection of objects (i.e., *data collection* in the figure). The IR system retrieves some

objects (e.g., documents) in response to the query. Finally, the user examines the retrieved objects. The existing query might be reformulated when more objects is desired. The process can be iterated until the user decides to stop the search.

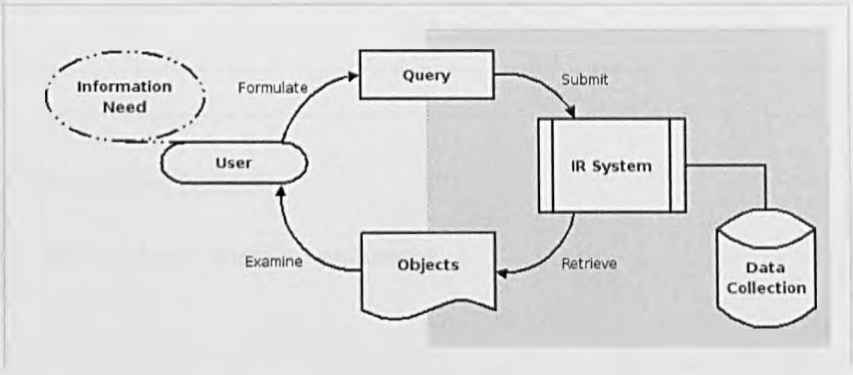



**Figure 2.1:** *A simplified view of information access process using an IR system*

In this chapter, we focus on the right hand side of the picture (i.e., the highlighted part). The other half of the picture is discussed in the following chapter. Note that the type of media being discussed in this thesis is mainly texts unless otherwise stated. The word *term* is used to refer to both single words and phrases.

## 2.2 Indexing

Indexing is a process of generating indexes for a collection of documents, just like the back of books. When a query is submitted to an IR system, the system looks up the indexes to retrieve a set of documents in which the query terms occur. Indexing allows IR systems to perform retrieval efficiently even with a large collection of documents. This section introduces several aspects involved in the indexing of a full-text collection.

**Sample documents** Let us think of two sample documents which are marked-up by an SGML scheme as shown in Figure 2.2. In this example, a whole document starts with <DOC> and ends with </DOC>. An ID of the document is given in the <DOCNO> and </DOCNO> tags. The content of the document is located between <TEXT> and </TEXT>.

```
<DOC>
<DOCNO>SMPL-001</DOCNO>
<TEXT>
This is the first sample document.
</TEXT>
</DOC>
```

```
<DOC>
<DOCNO>SMPL-002</DOCNO>
<TEXT>
This is the second sample document.
</TEXT>
</DOC>
```

**Figure 2.2:** *Sample documents*

**Preprocessing and tokenisation** The first process of indexing is to parse the documents to extract appropriate information from the texts. In our example, the sample documents can be parsed to extract the document IDs and contents. Once these pieces of information are extracted based on the SGML tags, the tokenisation can be applied to a relevant part of documents (e.g., content in <TEXT> and </TEXT>). The definition of tokens can vary, but for simplicity, a token is defined as any strings separated by non-word characters in our example. More discussion on the definition of tokens can be found in Grefenstette and Tapanainen (1994). The case of capital letters can also be lowered to simplify the indexing process. The result of the parsing and tokenisation is shown in Figure 2.3.

**Stopwords removal** A stopword is a word which appears in a document collection very frequently, thus, it is often regarded as little help to retrieve relevant documents. Due to the high frequency of appearance, the removal of stopwords also helps an IR system reduce the size of indexes. Articles such as *a*, *an*, and *the*, pronouns such as *it*, *this*, or *his*, and prepositions such as *of*, *in*, or *about* are the examples of stopwords. Table 2.1 shows some of the stopwords used in SMART system (Salton, 1971). The stopwords can be removed from the tokens, and the result can look like Figure 2.4.

| DOCNO | Token |
|---|---|
| SMPL-001 | this |
| | is |
| | the |
| | first |
| | sample |
| | document |
| SMPL-002 | this |
| | is |
| | the |
| | second |
| | sample |
| | document |

**Figure 2.3:** *A stream of tokens after the parsing and tokenisation*

**Table 2.1:** *Example of stopword (First 50 entries from SMART system (Salton, 1971))*

| | | | | |
|---|---|---|---|---|
| a | afterwards | already | any | appreciate |
| a's | again | also | anybody | appropriate |
| able | against | although | anyhow | are |
| about | ain't | always | anyone | aren't |
| above | all | am | anything | around |
| according | allow | among | anyway | as |
| accordingly | allows | amongst | anyways | aside |
| across | almost | an | anywhere | ask |
| actually | alone | and | apart | asking |
| after | along | another | appear | associated |

| DOCNO | Token |
|---|---|
| SMPL-001 | first |
| | sample |
| | document |
| SMPL-002 | second |
| | sample |
| | document |

**Figure 2.4:** *A stream of tokens after stopwords removal*

**Stemming** Stemming is an automatic removal of suffixes from words to normalise morphological variants. In our example, if a query is *samples* it would not match any document. However there may be a case where a user hopes to find a document in which the word *samples* occurs even when s/he submits the query *sample*. Apart from this singular/plural case, one might also hope to retrieve a document that contains a string *criminal organisations*, when a query contains *organised crime*. The stemming enables an IR system to find a document over the morphological variation of words. There are a number of approaches to stemming (Frakes, 1992). Two popular approaches, Porter's stemmer (Porter, 1980) and the inflectional stemming (Krovetz, 1993), are illustrated in Table 2.2.

**Table 2.2:** *Example of stemming*

| Word | Porter | Inflectional |
|---|---|---|
| organise | organis | organise |
| organised | organis | organised |
| organising | organis | organising |
| organisation | organis | organisation |
| organisations | organis | organisation |

As can be seen, Porter's stemmer generates the stem *organis* for all variants while the inflectional stemming only removes *s* from the plural form of *organisation*. As a result, while the output of the inflectional stemming is arguably more readable, Porter's stemmer would match more variants of a word compared to the inflectional stemming. The stream of tokens where Porter's stemmer has been applied is shown in Figure 2.5.

```
DOCNO      Token
--------   --------
SMPL-001   first
           sampl
           document
SMPL-002   second
           sampl
           document
```

**Figure 2.5:** *A stream of tokens after Porter's stemmer*

**Inverted files** The last stage of indexing is to create inverted files. An inverted file stores a link between words and documents. An example of inverted files is shown in Figure 2.6 based on the stream of tokens which have been illustrated so far. As can be seen, the inverted file stores a set of document for all words left in the token stream. Using this structure, an IR system can find a set of documents in response to a query efficiently. The information on the position of words in documents could also be stored in the inverted file which would enable an IR system to perform a proximity search for phrasal queries (not shown in our example).

| Word | DOCNO |
|---|---|
| document | SMPL-001 |
| | SMPL-002 |
| first | SMPL-001 |
| sampl | SMPL-001 |
| | SMPL-002 |
| second | SMPL-001 |

**Figure 2.6:** *Inverted file*

## 2.3 Ranking documents

The process of indexing has been illustrated using a simple example. It has also illustrated how an IR system could find a set of documents in response to a query using inverted files. This section discusses ranking methods of documents. Some of the major ranking models are introduced with basic concepts such as *term frequency* and *inverse document frequency*. See Chapter 2 and 13 in Baeza-Yates and Ribeiro-Neto (1999), to which I owe much discussion in the following two subsections, for a comprehensive review of various ranking techniques in IR.

### 2.3.1 Major ranking models

One of the characteristics of classic retrieval models is to use the frequency of occurrence of terms as the primary source of ranking methods. In other words, a term can be weighted differently based on the frequency of occurrence. There are two major factors to weight terms: term frequency, often denoted as TF, and inverse

document frequency, denoted as IDF (Sparck Jones, 1972). They can be described as follows.

Let $N$ be the total number of documents in a collection and $n_i$ be the number of documents in which an index term $t_i$ appears. Let $freq_{i,j}$ be the raw frequency of term $t_i$ in a document $d_j$ (i.e. the number of times $t_i$ appears in the text of $d_j$). Then, the normalised frequency, $f_{i,j}$, of $t_i$ in $d_j$ is given by

$$f_{i,j} = \frac{freq_{i,j}}{\max_l freq_{l,j}} \tag{2.1}$$

where $freq_{l,j}$ is the frequency of a term $t_l$ that has the maximum frequency computed over all terms which are mentioned in the text of $d_j$. If $t_i$ does not appear in $d_j$ then $f_{i,j} = 0$. In addition, the inverse document frequency of $t_i$, $idf_i$, can be given as follows.

$$idf_i = \log \frac{N}{n_i} \tag{2.2}$$

The notion of term frequency is relatively straightforward since it gives a greater weight when a term occurs in a document more frequently. The notion of inverse document frequency is described in Sparck Jones and Willett (1997) as follows.

> "The basis for IDF weighting is the observation that people tend to express their information needs using rather broadly defined, frequently occurring terms, whereas it is the more specific, i.e., low-frequency, terms that are likely to be of particular importance in identifying relevant material. This is because the number of documents relevant to query is generally small, and thus any frequently occurring terms must necessarily occur in many irrelevant documents; infrequently occurring query terms, conversely, have a greater probability of occurring in relevant documents and should thus be considered as being of greater potential importance when searching a database." (p. 307)

Using these two factors, a weight, $w_{i,j}$, of $t_i$ can be given in various ways. One of

them is a product of term frequency and inverse document frequency (i.e., TF*IDF) as shown below.

$$\begin{aligned} w_{i,j} &= f_{i,j} \times idf_i \\ &= \frac{freq_{i,j}}{\max_l freq_{l,j}} \times \log \frac{N}{n_i} \end{aligned} \tag{2.3}$$

The following weighting function has also been suggested by Salton and Buckley (1988) for query terms.

$$w_{i,q} = \left(0.5 + \frac{0.5 freq_{i,q}}{\max_l freq_{l,q}}\right) \times \log \frac{N}{n_i} \tag{2.4}$$

The TF, IDF, and TF*IDF are some of the most basic concepts in IR. They are used not only for ranking documents but also for query expansion which is discussed in the later chapter. The rest of this section introduces some major ranking models such as the vector space model and probabilistic model.

**Vector space model** Given that a weight is given for the terms in a query and matched documents, a vector space model is designed to compute the similarity between a query and document based on a $n$-dimensional vector space representation (Salton, 1968). A similarity of two vectors (i.e. $\vec{q}$ and $\vec{d_j}$) is quantified, for example, by the cosine of angle between the two vectors as follows.

$$\begin{aligned} sim(d_j, q) &= \frac{\vec{d_j} \bullet \vec{q}}{|\vec{d_j}| \times |\vec{q}|} \\ &= \frac{\sum_{i=1}^{n} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{n} w_{i,j}^2} \times \sqrt{\sum_{j=1}^{n} w_{i,q}^2}} \end{aligned} \tag{2.5}$$

where a query vector $\vec{q}$ is defined as $\vec{q} = (w_{1,q}, w_{2,q}, \ldots, w_{n,q})$ and document vector $\vec{d_j}$ as $\vec{d_j} = (w_{1,j}, w_{2,j}, \ldots, w_{n,j})$. The lesser the angle between the vectors is, the greater the similarity between a query and a document is. IR systems such as SMART (Salton, 1971) are based on a vector space model.

**Probabilistic model** A probabilistic model, on the other hand, is designed to measure the similarity between a query and documents by considering the probability of query terms occurring in relevant documents and non-relevant documents. For example, the similarity between them can be determined as follows.

$$sim(d_j, q) \sim \sum_{i \subseteq q} w_{i,q} \times w_{i,j} \times \left( \log \frac{P(t_i|R)}{1 - P(t_i|R)} + \log \frac{P(t_i|\overline{R})}{1 - P(t_i|\overline{R})} \right) \tag{2.6}$$

where $P(t_i|R)$ represents the probability of term $t_i$ occurring in relevant documents $R$, and $P(t_i|\overline{R})$ occurring in non-relevant documents $\overline{R}$. When the probability of documents being relevant is unknown for a query, it can be estimated as follows, for instance.

$$\begin{aligned} P(t_i|R) &= 0.5 \\ P(t_i|\overline{R}) &= \frac{n_i}{N} \end{aligned} \tag{2.7}$$

A well-known weighting scheme, *BM25* (Robertson et al., 1995b), takes into account other aspects such as document length, as shown below.

$$sim(d_j, q) \sim \sum_{i \subseteq q} idf_i \times \frac{(k_1 + 1) freq_{i,j}}{K + freq_{i,j}} \frac{(k_3 + 1) freq_{i,q}}{k_3 + freq_{i,q}} + k_2 |q| \frac{avdl - dl}{avdl + dl} \tag{2.8}$$

where $K$, $k_1$, $k_2$, and $k_3$ are collection-dependent constants. The $dl$ represents a document length, and $avdl$ represents an average document length across a collection.

Established IR systems such as Okapi (Robertson et al., 1995b, 1997) and InQuery (Allan et al., 1996; Callan et al., 1992) are based on a probabilistic model. More recently, language model approaches, which are an extension of a probabilistic model, have also been developed by researchers (e.g., Ponte and Croft, 1998). Some have reported a better performance than the classic models discussed above (Croft and Lafferty, 2003). Comprehensive reviews of probabilistic models can be found in Crestani et al. (1998) and Sparck Jones et al. (2000a,b).

**Other models (the web)** The retrieval models that have been discussed so far pursue a notion of similarity by using the frequency of occurrence of words as the main source. Search engines on the web are known to use other factors such as hyperlinks or click-through information to determine authority and popularity of documents.

A hyperlink is a link in a web page which points to an address of another web page using a form of HTML tags such as `<a href="http://www.url.com/">`. If a page has more in-links, the page is regarded as a greater authority than others, and can be ranked higher than fewer in-links documents. Brin and Page (1998) and Gibson et al. (1998) explored this idea to rank documents. A click-through is a count of users' click made on a record in a search result to access the contents of the record. If a particular page is clicked more frequently, then the page is regarded as having a greater popularity than the others have. However, in practice, major search engines are likely to use a combination of multiple evidences based on similarity, authority, and popularity. A review of methods and techniques used in the retrieval on the web is discussed in Kobayashi and Takeda (2000).

## 2.4 Systematic evaluation of the retrieval performance

This section discusses a systematic evaluation of retrieval performance based on IR test collections, and introduces some of the basic measurements frequently used in IR research. A user-centred evaluation of an IR system is discussed in a later chapter.

### 2.4.1 TREC test collections

A test collection in IR usually consists of three parts: topics, documents, and relevance judgements. Test collections provide a platform for researchers to evaluate and compare the retrieval effectiveness of IR systems based on the common ground. The Text REtrieval Conference (TREC) test collection is one of the major test collections, which is co-ordinated by the National Institute of Standards and Technology (NIST). TREC collections contain significantly larger volumes of document collections and a wide range of queries compared to earlier test collections such as the

Cranfield or CACM collection (Harman, 1992). This makes a testing of IR systems more realistic. TREC's document collections include different sources such as Wall Street Journal, Associated Press, San Jose Mercury News, Financial Times, and LA Times. Recent TREC tracks have between 2GB to over 400GB of web documents. Most documents are tagged by SGML as shown in Figure 2.7.

```
<doc>
<docno> WSJ880406-0090 </docno>
<h1> AT&T Unveils Services to Upgrade Phone Networks
Under Global Plan</h1>
<author> Janet Guyon (WSJ staff) </author>
<dateline> New York </dateline>
<text>
American Telephone & Telegraph Co. introduced the first
of a new generation...[omitted]
</text>
</doc>
```

**Figure 2.7:** *Example of TREC document (WSJ880406-0090)*

Information requests in TREC typically consist of a topic number, title, description, narrative, and they are similarly tagged by SGML as shown in Figure 2.8. TREC participants are allowed to use any parts of the requests in their systems unless a given task specifies otherwise.

```
<top>
<num> Number: 348
<title>Agoraphobia
<desc> Description: Is the fear of open or public places
(Agoraphobia) a widespread disorder or relatively unknown?
<narr> Narrative: Relevant documents contain data on this
physical/mental disorder, including information on the
person affected, profession of the individual, impact
on the life work of the individual, as well as any data
on how these individuals cope with this disorder.
</top>
```

**Figure 2.8:** *Example of information request in TREC (Topic Number 348)*

In TREC, a query can be formulated using the information request in an automatic or non-automatic way. The latter includes any manual intervention in a query construction. Participants are usually asked to submit the top 1000 documents for

each topic (e.g. AdHoc Track in TREC). A pool of documents is then generated by using a set of the top 100 documents submitted by every participant. Relevance of documents in the pool is assessed by either a topic generator or expert of the topic subject. Relevance judgements are recorded for each topic, and retrieval effectiveness is measured for participated systems. The following section discusses some of measurements used in IR evaluation.

### 2.4.2 Recall, precision, and other measurements

One of the main goals of IR systems is to retrieve a set of documents that are relevant to a user's information need. Recall and precision provide a different perspective of the number of relevant documents retrieved by an IR system. Let us illustrate them with two sets of documents: a set of documents that are relevant to an information need (i.e., Relevant set, $R$), and another set of documents that are retrieved by a system (i.e., Answer set, $A$). Using these two sets, recall and precision can be defined as follows:

$$\begin{aligned} Recall &= \frac{|R \cap A|}{|R|} \\ Precision &= \frac{|R \cap A|}{|A|} \end{aligned} \tag{2.9}$$

In other words, recall quantifies the portion of relevant documents retrieved by an IR system, while precision measures the portion of retrieved documents that are relevant. A precision-recall graph can then be generated by plotting a precision value at a set of recall levels. Typically, an 11-points scale $(0, 0.1, 0.2, \ldots, 1.0)$ is used for the recall levels where 1.0 represents the total number of documents retrieved by the system (e.g., 1000). Figure 2.9 shows an example of a precision-recall graph based on a real IR system.

Additional measurements have also been developed based on precision and recall, as follows.

**Precision at N docs** Precision at N docs shows the portion of relevant documents retrieved at the top N documents. Typically, N varies between 1 to 1000. For exam-

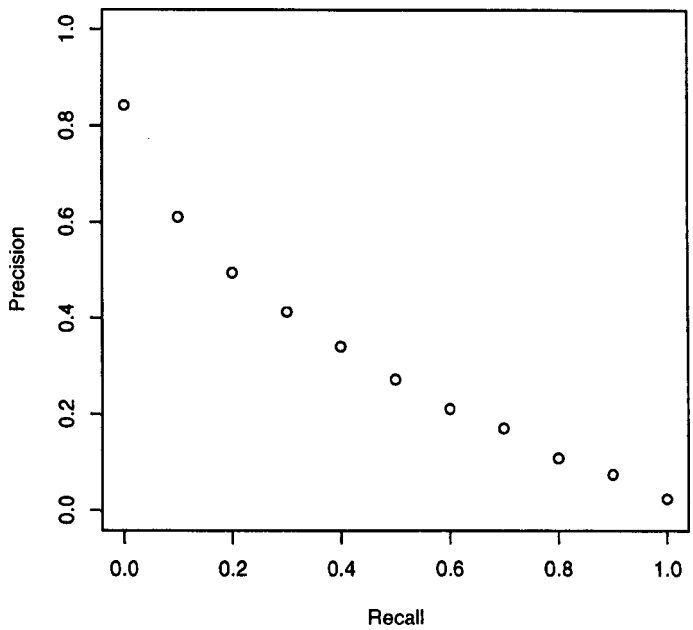



**Figure 2.9:** *Precision-recall graph*

ple, when there is one relevant document retrieved at the top 5 documents, precision at 5 docs is 0.2.

**R-precision** R-precision represents a precision at the $R$th retrieved documents where $R$ is the number of relevant documents known to a query (i.e., $|R|$). Therefore, when there are 10 relevant documents in a collection for a topic, and an IR system retrieves 8 relevant documents in the top 10 documents, then R-precision of the system is 0.8 for the topic.

**Mean average precision** Mean average precision (MAP) is commonly used to represent an overall performance of IR systems. Average precision is an average of precision obtained at the rank of each relevant document retrieved by a system. MAP is a mean value of average precision over the number of topics used in an experiment.

As can be seen, these measurements provide different perspectives of retrieval effectiveness in a systematic evaluation of IR systems. The measures can be used to understand how a change of parameters affects a system's performance, or to

compare more than one system under the same condition. While MAP is often used to indicate an overall performance of IR systems, precision at N docs can be used when the accuracy of retrieved documents at the top ranking is important in an experiment. This measure is also useful when the size of a Relevant set $R$ is unknown. R-precision can be used to analyse the retrieval performance based on a topic-by-topic comparison.

## 2.5 Summary

This chapter has introduced some of the basic concepts in IR. The stemming, stopwords, and inverted file structure were illustrated through a sample indexing process. The term weighting functions were discussed and major ranking models were introduced. Finally, a systematic evaluation of retrieval performance was discussed. Overall, this chapter has looked at information access process from a system-oriented perspective.

The next chapter starts to look at an interaction between an IR system and users. In particular, we review studies related to query reformulation.

# Chapter 3

# Towards interactive query reformulation

## 3.1 Introduction

In the paper "Helping people find what they don't know", Belkin (2000) illustrates some characteristics of the people who access to information systems, as follows.

> "When people engage in information-seeking behavior, it's usually because they are hoping to resolve some problem, or achieve some goal, for which their current state of knowledge is inadequate. This suggests they don't really know what might be useful for them, and therefore may not be able to specify the salient characteristics of potentially useful information objects.
>
> Unfortunately, typical information systems require users to specify what they want the system to retrieve. Furthermore, people engaging in large-scale information systems typically are unfamiliar with the underlying operations of the systems, the vocabularies the system use to describe the information objects in their databases, and even the nature of the database themselves." (p. 58)

Given that Belkin's view of information seekers can be applied to the majority of IR systems' users, it is perhaps not so surprising that formulating a query is a

difficult task. It is difficult in part because a query is supposed to adequately represent an underlying information need, and in part, because it is supposed to be well designed for a system and its resource, to retrieve relevant information successfully (Baeza-Yates and Ribeiro-Neto, 1999).

Consequently, improving a query representation initially given by a user has been one of the most active areas in IR. Such a process is called *query reformulation* (or query refinement). The majority of research on query reformulation is concerned with a technique called Query Expansion (QE). As mentioned in Chapter 1, QE refers to a technique which aims to supplement an existing query with additional terms (Efthimiadis, 1996). QE can also be seen as a process of finding candidate terms which can be used to better represent an underlying information need. While adding terms is just one form of query reformulation, short queries frequently submitted to IR systems are likely to benefit from such an augmentation.

This chapter provides a literature review of studies related to query reformulation. Firstly, methods and sources that have been used to derive a set of candidate terms for query expansion are surveyed. Secondly, query reformulation performed particularly in an interactive fashion using QE techniques is discussed. Finally, the limitations of existing methods and implications are addressed to propose a new approach to interactive query reformulation.

## 3.2 Methods and sources of query expansion

Approaches to query expansion are usually closely related to their sources of expansion terms. For example, Efthimiadis (1996) has broadly categorised QE techniques based on three types of sources: search results, collection-independent knowledge structures, and collection-dependent knowledge structure.

Search results are a set of documents retrieved by an IR system in response to a query. Collection-independent knowledge structures include manually constructed dictionaries or thesauri such as Collins' dictionary, Roget's thesaurus, and WordNet (Miller, 1995). They also include domain-specific classifications such as INSPEC, the ACM Computing Classification System, or Medical Subject Headings (MeSH). Collection-dependent knowledge structures, on the other hand, refer to

the structures which are automatically built upon a particular document collection. Efthimiadis' definition of the last category is quite broad and includes stemming, string similarity, as well as clustering and other types of structures.

Alternatively, Baeza-Yates and Ribeiro-Neto (1999) has grouped QE techniques into the approaches that use 1) feedback information from a user, 2) information derived from a set of documents initially retrieved (called a local set of documents), and finally, 3) global information derived from a document collection. The difference between Efthimiadis' categories and Baeza-Yates and Ribeiro-Neto's categories are not so significant, but some differences are found in the granularity and coverage of the sources. While Baeza-Yates and Ribeiro-Neto clearly distinguish a local set of documents from a user's feedback information derived from the local set, Efthimiadis does not. On the other hand, Efthimiadis' categorisation includes manually constructed knowledge structures which are excluded by Baeza-Yates and Ribeiro-Neto. This section discusses QE techniques by following the classification given by Baeza-Yates and Ribeiro-Neto (1999), but also covers the methods based on manually constructed knowledge structures.

### 3.2.1 Relevance feedback

A popular QE technique is based on relevance feedback (RF). As mentioned in Chapter 1, RF refers to a process of obtaining relevance information about retrieved documents. It typically either asks a user to identify a set of documents s/he think are relevant (Robertson et al., 1995b), or assumes that the top N ranked documents are relevant (Salton and Buckley, 1990). The basic idea of RF based QE is to update weight of terms based on relevance information. For example, weight of terms appeared in relevant documents can be increased and the top M highest terms can be selected as candidate expansion terms. There are a number of methods proposed as a term weighting function using RF. Let us take the Robertson/Sparck Jones F4 function (Robertson and Sparck Jones, 1976), for instance. In this model, $w$, weight of a term $t$ is represented as follows:

$$w = \log \frac{p(1-\overline{p})}{\overline{p}(1-p)} \tag{3.1}$$

where $p$ is the probability of $t$ occurring in relevant documents, and $\overline{p}$ is the probability of $t$ occurring in non-relevant documents. With the estimation of $p$ and $\overline{p}$ using the following functions,

$$\begin{aligned} p &= \frac{r}{R} \\ \overline{p} &= \frac{n-r}{N-R} \end{aligned} \tag{3.2}$$

the F4 function can be rewritten to:

$$w = \log \frac{r(N-n-R+r)}{(n-r)(R-r)} \tag{3.3}$$

where $N$ is the total number of documents in a collection, $n$ is the number of documents containing $t$, $R$ is the number of relevant documents for a topic, and $r$ is the number of relevant documents in which $t$ occurs.

In practice, a constant such as 0.5 is used in Equation 3.3 to alleviate the estimation of $p$ which is used in $w$ (see Sparck Jones et al. (2000a) for more detail), thus, it becomes:

$$w = \log \frac{(r+c)(N-n-R+r+c)}{(n-r+c)(R-r+c)} \tag{3.4}$$

Table 3.1 illustrates how the F4 function changes term weight over different conditions where a higher weight represents a greater level of significance. The size of a collection (i.e., $N$) is set to 10000, and constant $c$ is set to 0.5 in the table. As can be seen, the weight decreases as DF of a term increases (i.e., $n = 1$ to $n = 1000$). On the other hand, the weight increases as the number of relevant documents in which the term occurs increases (i.e., $r = 0$ to $r = 2$).

The term weight can be calculated for all terms in relevant documents. Then, the terms can be ranked in decreasing order of weight, and the top ranked terms can be used for expanding an existing query.

**Table 3.1:** *Term weight based on F4 function ($N = 10000$ and $c = 0.5$)*

| | | r = 0 | r = 1 | r = 2 |
|---|---|---|---|---|
| n = 1 | R = 0 | 3.82 | | |
| | R = 1 | 3.35 | 4.78 | |
| n = 10 | R = 0 | 2.98 | | |
| | R = 1 | 2.50 | 3.50 | |
| | R = 2 | 2.28 | 3.02 | 3.77 |
| n = 100 | R = 0 | 1.99 | | |
| | R = 1 | 1.52 | 2.47 | |
| | R = 2 | 1.29 | 2.00 | 2.70 |
| n = 1000 | R = 0 | 0.95 | | |
| | R = 1 | 0.48 | 1.43 | |
| | R = 2 | 0.25 | 0.95 | 1.65 |

In Harman (1988), terms are extracted from relevant documents and ranked based on three properties: posting, frequency, and noise. A posting is equivalent to $r$ in the F4 function. A frequency is a total term frequency of a term $t$ in relevant documents. The noise of $t$ is defined as follows:

$$noise = \sum_{i \subseteq N} \frac{freq_i}{freq_c} \times \log_2 \frac{freq_c}{freq_i} \tag{3.5}$$

where $N$ is the total number of documents in a collection, $freq_i$ is term frequency of $t$ in a document $d_i$, and $freq_c$ is a total term frequency of $t$ in an entire collection. Table 3.2 illustrates how the noise function varies across different conditions ($freq_c = 10000$). As can be seen, the noise increases as $freq_i$ increases.

**Table 3.2:** *Term weight based on Harman's noise function ($freq_c = 10000$)*

| $freq_i$ | $\frac{freq_i}{freq_c}$ | $\log \frac{freq_c}{freq_i}$ | noise |
|---|---|---|---|
| 1 | 0.000 | 13.288 | 0.001 |
| 2 | 0.000 | 12.288 | 0.003 |
| 5 | 0.001 | 10.966 | 0.006 |
| 10 | 0.001 | 9.966 | 0.010 |
| 20 | 0.002 | 8.966 | 0.018 |
| 50 | 0.005 | 7.644 | 0.038 |
| 100 | 0.010 | 6.644 | 0.066 |
| 1000 | 0.100 | 3.322 | 0.332 |

Among several ways of combining the three properties, Harman's experiment showed that a product of three (i.e., *posting* × *frequency* × *noise*) performed best for ranking expansion terms. When the top 20 terms ranked by the best method were used, the expanded queries improved MAP by 9.4% over initial queries. This improvement was based on relevant information of the top 10 retrieved documents. Harman also investigated if further improvement can be achieved by examining an increased number of retrieved documents for expansion term selection. More specifically, Harman filtered the original top 20 terms by examining whether or not they appeared in the relevant documents found in the top 20 retrieved documents. The filtering reduced the average size of expanded queries to 12 terms but improved MAP by 31%.

Another experiment of term weighting algorithms using RF based QE was carried out by Efthimiadis (1993). The experiment focused on a user's term selection as opposed to retrieval effectiveness of expanded queries. Efthimiadis compared six term weighting algorithms by analysing a distribution of the terms which subjects thought to be useful for search. Table 3.3 summarises the algorithms tested in Efthimiadis' experiments.

**Table 3.3:** *Term weighting functions tested in Efthimiadis (1993)*

| Name | Function |
|---|---|
| F4 | $\log \frac{(r+0.5)(N-n-R+r+0.5)}{(n-r+0.5)(R-r+0.5)}$ |
| F4 modified | $\log \frac{(r+c)(N-n-R+r+1-c)}{(n-r+c)(R-r+1-c)}$ where $c = n/N$ |
| MUSCAT | $\frac{r}{R} - \frac{n}{N}$ |
| EMIM | $\log \frac{rN}{Rn} r - \log \frac{(n-r)N}{(N-R)n}(n - r) - \log \frac{(R-r)N}{(N-n)R}(R - r) +$ $\log \frac{(N-n-R+r)}{(N-n)(N-R)}(N - n - R + r)$ |
| $w(p-q)$ | $\log \frac{(r+0.5)(N-n-R+r+0.5)}{(n-r+0.5)(R-r+0.5)}(\frac{r}{R} - \frac{n-r}{N-R})$ |
| ZOOM | $n$ (within the top 50 documents) |

$N$ is the total number of documents in the collection, $n$ is the number of documents containing the term $t$, $R$ is the number of known relevant documents, and $r$ is the number of known relevant documents in which the term $t$ occurs.

To briefly summarise the tested algorithms, *F4* is the Robertson/Sparck Jones F4 function which has previously been discussed. *F4 modified* is a variant of F4 function where a constant $c$ is set to $\frac{n}{N}$. The *MUSCAT* algorithm proposed by Porter and Galpin (1988) is defined by the portion of relevant documents in retrieved documents and a variant of inverse document frequency. The *expected mutual information*

*measure* (EMIM) is discussed in Van Rijsbergen (1979) and uses a two-by-two contingency table that consists of the presence or absence of a term in a document, and relevance or non-relevance of the document. The $w(p-q)$ is another variant of the F4 function with an additional adjustment based on the difference between the probability of a term being in relevant documents and non-relevant documents. *Zoom* was based on the term frequency of the top 50 ranked documents.

Relevance feedback information was collected from the transaction log of 25 real users who searched upon the INSPEC database. Candidate expansion terms were ranked based on the RF, and the top five terms were selected from each algorithm. A pool of the expansion terms was presented to subjects who judged the relevance of the terms. Efthimiadis' experiments show that the algorithms, which are based on a different theoretical motivation, nonetheless generate a similar ranking of terms. Specifically, EMIM and $w(p-q)$, F4 and F4 modified, and MUSCAT and ZOOM were found to produce similar results. Consequently, the distribution of users' selection was also found to be similar in these pairs. However, the first group (i.e. EMIM and $w(p-q)$) retrieved the most terms which subjects thought to be useful.

As can be seen, RF has been extensively studied by researchers as a means of QE. The source of expansion terms was mainly a user's relevance judgement on a set of retrieved documents. The next section discusses techniques that are based on an analysis of an entire document collection.

### 3.2.2 Global analysis

Global analysis refers to an analysis of an entire document collection, as opposed to a particular set of documents. One of the applications based on the global analysis is term clustering. Term clustering is a statistical technique that groups together similar terms in a multidimensional space. As mentioned in Chapter 1, term clustering is often motivated by the association hypothesis, which states "if an index term is good at discriminating relevant from non-relevant documents, then any closely associated index terms is also likely to be good at this" (p.134) (Van Rijsbergen, 1979). The similarity between terms can be measured by various properties. An example

is based on the number of documents in which a pair of terms co-occurs.

Applications of term clustering have a variety of names such as the term-term relationship matrices (Salton, 1980), co-occurrence list (Schatz et al., 1996), co-occurrence based thesaurus (Schütze and Pedersen, 1997), associative thesaurus (Sebastiani, 2001), similarity thesauri (Qiu and Frei, 1993), and so forth. A region of documents for counting the co-occurrence of terms can be an entire document or a part of the document. Also, the co-occurrence list can be generated by a comparison of pairs of terms or comparison of neighbours of terms. The former is called the first order co-occurrence list and the latter is called the second order list (Qiu and Frei, 1993).

A process of generating a term cluster is described by Sebastiani (2001) as follows.

1. standard (automatic) document indexing is performed, thereby generating the usual term-document incidence matrix that specifies a weight $w_{i,j}$ for each pair $< t_i, d_j >$ constituted by a term $t_i$ and a document $d_j$;

2. a term-term matrix is generated, specifying a *semantic relatedness* value $r_{k,m}$ for each pair of terms $< t_k, t_m >$. The matrix may be symmetric or not, depending on the underlying notion of similarity, and is generated by taking into account factors such as degree of co-occurrence or co-absence of the two terms in the collection;

3. small values of $r_{k,m}$ are replaced by 0, thus leaving only highly related pairs with a nonzero coefficient;

4. these latter, which determine the edges of the resulting associative thesaurus, may be identified:

   (a) as the top $n$ coefficients for any given term, for a pre-specified value of $n$;

   (b) as all the coefficients exceeding a pre-specified threshold value.

Having the basic steps in mind, the main issue of term clustering is how to determine the semantic relatedness, $r_{k,m}$. While early studies used the first-order co-occurrence data (Salton, 1980), the second order co-occurrence has also been exploited by Schütze and Pedersen (1997). The second order co-occurrence is a measure based on the neighbours of $t_k$ and $t_m$, as opposed to their co-occurrence itself.

Grefenstette (1992) also applied the second order co-occurrence analysis but only for a particular set of parts-of-speech that can be seen as a modifier of words.

A limitation of term clustering based on an entire collection was pointed out by Peat and Willet (1991). They illustrates that terms in a cluster are often found in the documents that have already been retrieved by query terms, thus, clustered terms' contribution to the improvement of retrieval effectiveness (recall, in particular) tends to be limited. Also, frequent words are likely to belong to many clusters which could harm the retrieval performance.

### 3.2.3 Local analysis

Another approach to generating a term cluster is based on a set of retrieved documents. An example is called *local context analysis* (LCA). LCA aims to analyse the top $N$ documents retrieved in response to a query (Xu and Croft, 1996). The major differences from the previously discussed term weighting functions are that it uses the top ranked passages as opposed to documents, and that it uses noun phrases as opposed to all words. A passage can be defined as a text window of fixed number of words.

In LCA, the weight, *bel*, of a concept (noun phrase), $c$, that is found in the top $N$ passages retrieved by a query, $Q$, is defined as follows.

$$bel(Q,c) = \prod_{t_i \in Q} \left( \delta + \log(af(c,t_i)) \cdot \frac{IDF_c}{\log(n)} \right)^{IDF_i} \tag{3.6}$$

where

$$af(c,t_i) = \sum_{i=1}^{j=n} tf_{i,j} \cdot tf_{c,j}$$

$$IDF_i = \max\left(1.0, \log_{10}(N/N_i)/5.0\right)$$

$$IDF_c = \max\left(1.0, \log_{10}(N/N_c)/5.0\right)$$

where $tf_{i,j}$ is the number of occurrences of a query term, $t_i$ in a passage, $p_j$, $tf_{i,c}$ is the number of occurrences of $c$ in $p_j$, $N$ is the number of passages in a collection, $N_i$ is the number of passages in which $t_i$ occurs, $N_c$ is the number of passages in which

**Table 3.4:** *Boolean expressions generated by a search-aid thesaurus (INT: initial query term, SYN: synonymous, NT: narrower term, RT: related term)*

| Search type | Boolean Expression |
|---|---|
| Synonym search | ( $INT_i$ **OR** $SYN_i$ ) **AND** ( $INT_j$ **OR** $SYN_j$) |
| Narrower search | ( $INT_i$ **OR** $NT_i$ ) **AND** ( $INT_j$ **OR** $NT_j$ ) |
| Related term search | ( $INT_i$ **OR** $RT_i$ ) **AND** ( $INT_j$ **OR** $RT_j$ ) |
| Union search | ( $INT_i$ **OR** $SYN_i$ **OR** $NT_i$ **OR** $RT_i$ ) **AND** ( $INT_j$ **OR** $SYN_j$ **OR** $NT_j$ **OR** $RT_i$) |

$c$ occurs, and finally, $\delta$ is a constant to avoid *bel* from being zero. $\delta$ is set to 0.1 in Xu and Croft (1996). In this formula, the $af(c, t_i)$ part aims to reward concepts based on the co-occurrence with query terms. Two IDFs are used to penalise frequently occurring concepts and rewards infrequently ones. The IDFs are defined such that any occurrences lower than 1/100,000 of the collection get the score of 1.0.

They compared the performance of LCA to a pseudo RF technique which assumed the top N documents were relevant in query expansion. The result based on the TREC-4 collection show that an overall performance of LCA is slightly better than the pseudo RF technique. However, LCA improved more topics and degraded fewer topics than the pseudo RF, which suggests that LCA can be more robust than a pseudo RF technique. Later, additional experiments were carried out and presented in Xu and Croft (2000). The experiments show that, although they assumed that a passage should contain less noise than a document, thus, using a passage should provide a better result, no significant difference was found between a passage-based LCA and document-based LCA in retrieval effectiveness.

### 3.2.4 Manually constructed structures

The techniques discussed so far are based on a statistical analysis of retrieved documents and/or entire document collection. An alternative approach to QE is to use semantically related terms that are defined by a manually constructed thesaurus (Greenberg, 2001b; Kekäläinen and Järvelin, 1998; Kristensen, 1993; Voorhees, 1994).

For example, QE can be carried out by using semantically related terms (i.e., narrower, border, synonyms, or related terms) with Boolean operators. Kristensen (1993) devised four types of expansion query as illustrated in Table 3.4. In their

approach, for each term in an initial query, related terms are extracted from a thesaurus, a structured query was built using the OR operator to connect the initial query term and thesaurus terms, and using the AND operator to connect the OR pairs. As can be seen, the Union search subsumes the other types in terms of recall. A union search was designed to use all types of related terms in an expanded query.

Their experiments used two sets of topics. The first set consists of 30 topics and was used to compare the baseline search (with no expansion) to the Union search. The second set consists of 14 topics which was a subset of the first set. The second set was used to compare all types of searches. They created a Finnish test collection that consisted of 227,000 news articles, with a total of 1384 relevance judgements for 30 topics. Their experiments show that the contributions of different relations are:

- For recall: Union > SYN & NT > RT
- For precision: NT > RT > SYN > Union

A similar experiment was carried out by Greenberg (2001a) using an English thesaurus called ProQuest (ProQuest, 1997) and ABI/Inform document collection via DIALOG. Greenberg's study included Border Terms (BT) and an expanded query was constructed in a similar way to Kristensen (1993). The result of Greenberg's experiment was:

- For recall: RT & BT > NT > SYN
- For precision: SYN > NT > BT > RT

Therefore, the two studies show a slightly different order with regard to what relation improves recall and precision. However, both studies show that the Boolean operator which improves the precision does not help improve recall, and vice versa. This suggests that a particular relation is unlikely to be useful for query expansion across topics. Voorhees (1994) also expressed a difficulty in selecting an appropriate relation to add to initial queries. Voorhees states that "the most useful relations for query expansion are idiosyncratic to the particular query in the context of the particular document collection." With the larger first set's experiment in Kristensen (1993), the Union search reduced precision by 10% (51.2% as opposed to 62.5% of

the baseline) while recall was more than doubled. Therefore, although the improvement can vary over the types of relations, adding all relations in an union search appears to be a reasonable remedy for this problem.

## 3.3 User interaction with candidate terms

The previous sections discussed the sources and techniques of QE. The following sections discuss some aspects of human-computer interaction (HCI) in the context of QE. QE performed in an interactive fashion is called interactive query expansion (IQE). Automatic query expansion (AQE), on the other hand, refers to QE performed automatically. A typical scenario of IQE is that an IR system suggests a set of candidate expansion terms to a user who decides which terms in the suggestion to add to an existing query.

Since IQE is an interactive activity between a user and IR system, users' information seeking behaviour is also of our interest. However, since experiments with real users are expensive, there have been a number of experiments that *simulate* users. In this section, we first review studies which are based on the user simulation, followed by studies which investigate real users. We also discuss the presentation of expansion terms.

### 3.3.1 Simulating the interaction

**Harman's simulation** The simulation of real users is typically carried out by using a test collection which has a record of relevant documents judged by subject experts in response to a set of topics. For example, Harman (1988) used the Cranfield collection to evaluate her noise model (See Section 3.2.1). Relevant documents found in the top 10 ranked documents were used under a simulated situation where a user was assumed to make a correct judgement on the documents.

Harman also investigated if further improvement could be achieved by a user's additional effort of making a relevance judgement on lower ranked documents. Harman selected expansion terms only when they also appeared in at least one of unseen relevant documents (i.e., relevant document ranked lower than the top 10).

The experiment shows that the filtering reduced the average length of expanded queries, yet contributed to a substantial improvement in MAP. While whether or not a real user is willing to make such effort is open for discussion, Harman's experiment suggests that additional user effort can lead to a better term selection. It also suggests that a fixed length of expanded queries may not achieve the best performance.

**Magennis and van Rijsbergen's simulation** Motivated by Harman's work, Magennis and van Rijsbergen (1997) introduced the size of expansion terms as a variable in a simulation of *experienced users*. Their experiment also considered the effect of the iterative searches which were more likely to happen in a real situation. Their method of simulating an experienced user is as follows.

- Run an initial search, and the top 20 ranked documents are used for query expansion;
- On each query expansion iteration, rank the terms in the 20 documents based on the Robertson/Sparck Jones F4 function score that are applied to previously unretrieved relevant documents;
- Select the highest ranked terms by applying a cut-off to this ranking and add those to the query;
- Run a search using every possible combination, over four query expansion iterations, of the cut-off values 0, 3, 6, 10, and 20 to generate expanded queries;
- The best retrieval effectiveness is used as an experienced user's performance.

Therefore, this simulated an experienced user who would select the best number of expansion terms in each iteration of search.

The simulation of an experienced user was compared to the performance of an AQE using the TREC-3 AdHoc Track collection. Their AQE was based on the F4 function, and two cut-off values (6 and 20) for the size of expansion terms were tested. The comparison between the experienced user and AQE was made by precision at 100 documents after four iterations of the search. The result shows that the

performance of the simulated IQE was consistently better than the AQE except for one topic. This suggests that, when the best number of expansion terms is selected by a user, IQE can be more effective than AQE.

However, their subsequent experiment with five real users shows that this performance can be difficult to achieve by an inexperienced user. Five subjects were asked to select terms from the 20 suggested terms that were identical to the AQE set. After four iterations of relevance judgement and term selection, the performance was compared to the AQE with the cut-off of 6. Precision was improved for 8 topics, degraded for 10 topics, and no change for the remaining 2 topics. However, the difference was not significant for most topics. These findings led the authors to conclude that:

- AQE using RF terms offers a large overall improvement, but the improvement is very variable across topics;
- IQE, as it might be performed by experienced users, offers a small but significant further improvement. The improvement is very consistent across a range of topics;
- Inexperienced users of IQE did not make good term selections and failed to improve on AQE. The lack of improvement is also consistent.

The simulations show an advantage in research such that one can test a number of different parameters of their system without involving real users. For example, a total of 400 runs (20 topics * 4 iterations * 5 cut-offs) were tested by Magennis and van Rijsbergen (1997), which would be infeasible to carry out with real users.

**Ruthven's simulation** More recently, Ruthven (2003) carried out a comprehensive analysis of potential effectiveness of RF based IQE. Ruthven's simulation was based on 32,678 possible combinations of expansion terms, that were derived from the top 15 candidate terms for each query. The candidate terms were extracted from the top 25 retrieved documents of TF*IDF based initial searches, and ranked by $w(p-q)$ function (See Table 3.3 on page 25). Ruthven's work tested every possible length of expansion terms up to 15, and used three document collections: Associated Press

(AP), San Jose Mercury News (SJM), and Wall Street Journal (WSJ). However, the factor of iteration was not considered in the analysis.

Another difference from previous simulated studies was that two groups of topics were excluded from the experiments. One was those topics which did not contain a single relevant document in the top 25 retrieved documents by initial queries. Another was the topics where all relevant documents were retrieved by the initial queries. The two groups were excluded since they would not make any changes to initial results. As a result, a total of 51 queries were excluded from the original 150 queries.

Ruthven's experiments show that it is possible to improve retrieval effectiveness in more than 94% of topics of the three collections, when the best combination of expansion terms is used. The average improvement in MAP was 22.3% (AP), 29.1% (SJM), and 22.4% (WSJ), respectively. However, it also shows that a middle-performing query, which is 16384th of all combinations, can only improve between 30% and 38% of queries.

Ruthven also analysed the relationship between the performance and two factors: query length and $w(p-q)$ scores. It showed that the average size of expanded queries was very similar across the best, middle, and poor IQE performance, and the length was somewhere between 7.16 and 7.61, except 5.56 in the best combination for the SJM collection. This suggests that giving extra terms does not necessarily gain an improvement in retrieval effectiveness. The relationship between the performance and $w(p-q)$ score was more evident. In other words, the best combination tended to have a higher $w(p-q)$ score than the middle and poor performing queries (e.g., 2.94 in the best and 1.92 in the middle of AP). This suggests that a $w(p-q)$ score can be more indicative than a query length regarding the potential improvement of retrieval effectiveness.

As can be seen, the simulation of user actions has provided further insights into the potential effectiveness of IQE, which would have been difficult to obtain with real users. The comparison with real users' term selection also suggests that simply converting an AQE technique into an interactive fashion does not always work, thus, further investigation should be carried out to improve the performance of IQE. The next section looks into how users' information seeking behaviour is

related to IQE.

### 3.3.2 Information seeking behaviour

The previous section highlighted the characteristics of IQE compared to AQE based on simulated users. Another important factor in the investigation of IQE is the interaction of real uses with an IQE facility. This section discusses some aspects of human computer interaction (HCI) in the context of IQE where real users are often involved in the studies.

**Take-up rate, cognitive load, and user control** One of the fundamental questions in IQE is whether a user of an IR system is willing to use an IQE facility to complete a search task. For example, a series of experiments carried out by Hancock-Beaulieu et al. (1995) measured a take up rate of suggested terms in a live operational setting with real users and their real information needs. A RF based IQE facility was implemented in a library catalogue database in City University Library using an X windows system. When relevance feedback was given, terms in relevant documents were ranked and the top 24 terms were presented to a user.

In the experiments, a total of 591 searches were recorded during a period of two months. Out of 591 searches, 383 (65%) were not reformulated, 144 (24%) were manually reformulated by users, and 58 (9.8%) were reformulated interactively by using their IQE facility. The authors commented that they felt the take up rate was lower than expected. A subsequent experiment with a different interface showed a slightly higher take-up rate (Beaulieu, 1997). This suggests that users are not always willing to engage with an IQE facility. However, it appears that the interface design can change the accessibility of an IQE facility.

Hancock-Beaulieu et al. (1995) also reported that "half of the subjects declared that they had experienced some problems in determining the appropriateness of terms", and they speculated that "the display and selection of individual terms from a list also may have made it difficult for users to consider the expanded query as an entity, or to establish a relationship between the different term". This suggests that the presentation of candidate terms can be an important factor for a successful use of an IQE facility.

Further investigation was undertaken by Beaulieu:97 looking at users' cognitive load related to the interface and functionality of IR systems. Beaulieu suggested that the cognitive load appeared to depend on two interrelated factors: (a) to what extent the functionality of a system's operation can be made visible; and (b) how control of the interaction can be distributed between a user and system.

The work by Shiri and Revie (2003) provides some insight into the first factor. They investigated two types of moves in user interaction: cognitive moves and physical moves. The first type was defined as a move in which users performed some kind of conceptual analysis of terms or documents. Some of the cognitive moves were term input, browsing terms in the thesaurus, selection of terms, combine search terms, browsing retrieved titles, and query reformulation. The second type was defined as a move that was associated with the use of system features. Some of the physical moves were perform search, scroll up and down, a browser's back and forward, and previous and next pages of search results.

Their analysis of the moves was made in the light of the complexity and familiarity of topics. The complexity of topics was determined by the number of search terms and Boolean operators. A topic was classified as a simple topic when the search terms were less than three or the operators were less than two. The familiarity of topics was determined by participants' subjective assessments based on a selection of Unfamiliar, Moderately familiar, and Very familiar.

A thesaurus enhanced database in the areas of Agricultural Science was used in their experiment. The system allowed users to map their search terms to the thesaurus. 30 subjects were recruited and carried out the searches for their own information needs (3 topics). The moves during the searches were recorded and analysed for a total of 90 topics. Of those 90 topics, 60% were classified as a simple topic and the rest as a complex topic. As for the familiarity, 55% were judged as Moderately familiar, 25% as Very familiar, and 20% as Unfamiliar.

Their analysis shows that the complexity of topics affects the number of moves more significantly than the familiarity. While the simple topics involved on average 9.9 cognitive moves and 17.3 physical moves, the complex topics involved on average 16.2 and 24.7, respectively. As for the familiarity, unfamiliar topics involved on average 10.8 cognitive moves while Moderately and Very familiar topics involved

13.0 and 12.3 moves, respectively. A similar trend was found in the physical moves. These results show that users require a high level of engagement and effort in complex topics, which supports the finding of Fowkes and Beaulieu (2000). Also, the results suggest that a greater degree of support should be provided for the topics that are unfamiliar to users.

Koenemann and Belkin (1996) investigated the second factor, that is, to what extent the user control can be beneficial in query expansion. Four different degrees of the user control on query expansion were devised as follows:

- **Baseline**: No relevance judgement, no term presentation, and no term selection by users.
- **Opaque**: Relevance judgement only. No term presentation and no term selection.
- **Transparent**: Relevance judgement and term presentation. No term selection.
- **Penetrable**: Relevance judgement, term presentation, and term selection by users.

For example, in the Transparent condition, when a user made a relevance judgement, the system presented a set of expansion terms which were added to the subsequent search, but they were not allowed to change the expansion terms. However, a user was allowed to modify an existing query manually. In the penetrable condition, on the other hand, a user was allowed to select terms from a list of suggested terms.

Relevance feedback and query expansion were implemented using a probabilistic IR system, InQuery (Callan et al., 1992). The interface design was similar to the one used in Beaulieu (1997). Two search topics (162 - Automobile Recalls and 165 - Tobacco Advertising and the Young) were selected from the TREC topics along with the WSJ document collection.

64 subjects were recruited for their experiments. Every subject was asked to perform searches on the first topic with the baseline, followed by the second topic with

one of the three conditions (i.e. Opaque, Transparent, and Penetrable). Twenty minutes were given for each topic. Retrieval effectiveness was measured by precision at 30 documents.

The results show that the average precision of the final queries in the RF conditions is 17 to 34% higher than the baseline condition. The Penetrable condition showed the best performance among the four conditions, being 15% better than the Opaque and Transparent conditions. The average number of expansion terms generated by the system was between 17.8 and 42.8 across the RF conditions. The subjects manually generated a mean of 3.8 terms. Therefore, most terms in the queries were derived from the query expansion performed by the system. The average size of queries in the Penetrable condition, however, was 10 to 20 terms fewer than the other two RF conditions, suggesting that the subjects made a few selection on the suggested terms. The average number of iterations in the Penetrable condition was also lower than the other conditions.

The results suggest that a high degree of control given to a user can be beneficial in query expansion. The authors reported that most subjects preferred relevance feedback to the baseline search, and some expressed their desire to "see and control". This appears to be in contrast with the observations made by Beaulieu (1997). There are two major differences between the two studies: topics and training time. Koenemann and Belkin (1996) used two pre-defined topics with the detail descriptions about the relevance, while Beaulieu obtained the data from a live system which could contains various search topics. Furthermore, the subjects in Koenemann were given up to 40 minutes of training with the system while it is our understanding that no explicit training was given in Beaulieu's experiment.

**Semantic relations of expansion terms** Another body of research on HCI aspects of IQE examines the semantic relationships between initial query terms and expansion terms that are selected by users (Anick, 2003; Beaulieu, 1997; Efthimiadis, 2000; Greenberg, 2001b; Jones et al., 1995). Studies discussed in this section often use semantic relations defined by a thesaurus.

Jones et al. (1995) used the INSPEC thesaurus and associated documents that were indexed by the Okapi system. The system listed thesaurus terms which com-

pletely or partially matched query terms, ranked by a variant of TF*IDF score. On average, 148 terms were presented to 39 subjects as candidate expansion terms. Of those, 80 terms included a query term, 20 terms were located in a single distance (i.e. one node) from the query terms in the hierarchy, and 25 terms were located in the distance of two nodes. The breakdown of the semantic relations showed that more than half of suggested terms were related terms (RTs) of query terms, nearly 10% was the narrower terms (NTs,) and 6% was the broader terms (BTs). On average, 10% of suggested terms were evenly selected by users from the three relations. Although an overall performance of the expanded queries was not significantly better than the initial queries, Jones et al. (1995) commented that the performance was generally better when a large number of terms were available in the thesaurus.

Efthimiadis (2000) carried out another user study in a RF based setting using the INSPEC collection. In this study, the expansion terms were extracted from perceived relevant documents as opposed to a thesaurus. Efthimiadis' study with 25 subjects reported a higher rate of term selection than that of Jones et al. (1995), and a mean of 28% of terms were selected. The number of suggested terms varied from 34 to 137 across the subjects by users. The breakdown of the semantic relations between initial query terms and selected expansion terms were that 46% was of NTs, 34% had no relation defined in the thesaurus, 17% was of RTs, and 3% was of BTs. Efthimiadis also examined the subjects' perceptions about the association of their selection in multiple choices. 88% of selected terms were regarded as RTs by the subjects, 66% were thought as synonyms. This suggests that the subjects are not necessarily aware of the correct semantic relations of expansion terms. Furthermore, 34% of expanded terms were thought to be representing new ideas, as opposed to have some semantic association.

The result showed that retrieval effectiveness was improved by the expanded queries. While initial queries retrieved on average three highly relevant documents, the expanded queries retrieved on average further nine highly relevant documents. Due to the operational system used in his study, the number of documents examined was ranged from 16 to 66 with an average of 33 documents per topic. The improvement on retrieval effectiveness appears to be in contrast with the result of Jones et al. (1995). A reason for the difference between the two studies could be the

source of expansion terms. Efthimiadis extracted terms from relevant documents while Jones et al. selected terms from the thesaurus. In Efthimiadis' study, more than a third of selected terms did not have a semantic relation with query terms in the thesaurus. This suggests that the selection of expansion terms should not be limited to those that are associated in a thesaurus.

Anick (2003) has extended the thesaurus-like relationships and added the functional relationships such as Head, Modifier, Elaboration, and Location in the investigation. An example of Head category was *Triassic period* as an expansion term of *Triassic*. Similarly, an example of Modifier was *plastic buckets* as an expansion term of *buckets wholesale*. The Elaboration was defined as "a term which restricts or adds further context to the concept expressed in the original query". The analysis based on a set of 100 query expansion sampled from a search engine's query log, showed that the largest categories were Head, Modifier, and Elaboration which in total comprised over 65% of the sample set. The three categories can be seen as a variant of NTs in a traditional thesaurus classification. However, this also demonstrates that the syntactic aspect of the relationships between an initial query and expansion terms can be used to understand the information seeking behaviour of users in interactive query reformulation.

### 3.3.3 Presentation of expansion terms

As discussed above, the design of a search interface can be a factor to affect information seeking behaviour. In interactive query reformulation, therefore, it is possible that a different way of presenting expansion terms can affect how users interact with a term suggestion facility. However, studies appear to be limited in this area.

Lists have been used as the form of presentation of expansion terms in the vast majority of research on IQE (e.g. Harman, 1988; Allan et al., 1996; Beaulieu, 1997; Robertson et al., 1997; Efthimiadis, 2000; Belkin et al., 1999, 2003). This is true even when a structured knowledge resource such as a thesaurus is used as the source of expansion terms (Greenberg, 2001a; Jones et al., 1995). Lists are simple for a system developer to implement and can be intuitive for a user to interact. Since most techniques in QE involve some sort of *ranking* of candidate expansion terms, the top N

ranked terms can be displayed in a list. Typically, 6 to 30 terms are presented to users. The experiment of Harman (1988) is an early example of presenting expansion terms in a list. A user is then asked to select the terms of interest from the list to add to an existing query. This interactive approach has been a standard way of IQE and adopted by many studies (See Figure 3.1 for example).

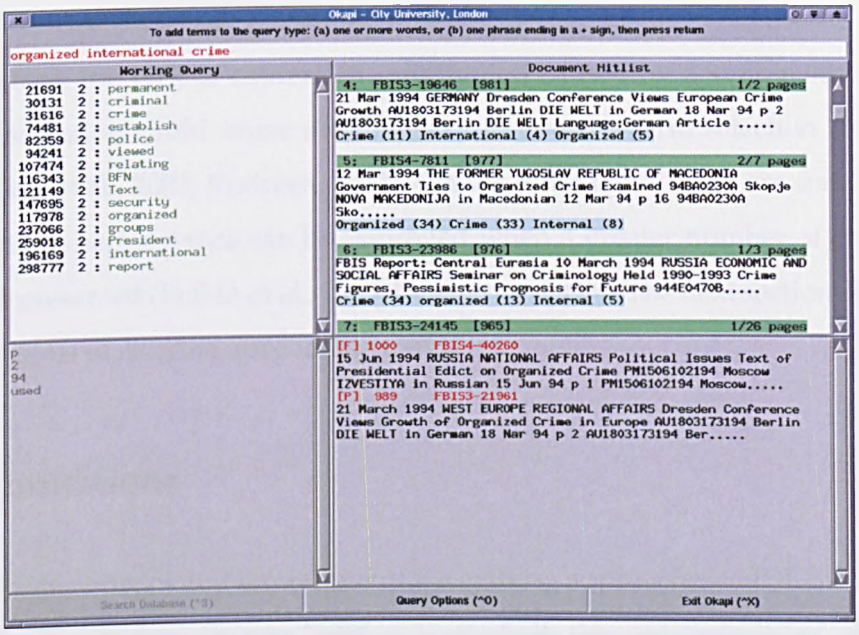



**Figure 3.1:** *A ranking-based list form of presentation of expansion terms in Okapi (taken from Beaulieu and Gatford (1998))*

Alternatively, tables were investigated by Anick (2003) where twelve candidate terms were displayed in a four by three matrix table. Two methods of ordering candidate terms were tested. One was an alphabetical order and another was a category-based method. The latter listed four terms in three categories: phrases containing a query term, other multi-word phrases, and single word terms. The experiment shows that the presentation order can affect a user's term selection. For example, while the selection in the alphabetical order was almost evenly distributed across different positions in the table, the category order had a distribution skewed towards a particular position in the table.

A disadvantage of simple structures such as a list appears to be the lack of contextual information on suggested terms. For instance, the literature has reported that users found it difficult to select appropriate terms for QE when:

- the terms are either too specific or unfamiliar;

- it is not clear why and how the terms are selected and ranked;
- it is not clear how the terms are associated with initial query terms;
- it is not clear how the terms are related each other; and/or
- it is not clear how the terms are used in the texts.

Therefore, the lack of contextual information caused by a simple structure of term presentation could cause the difficulty in a user's term selection (Beaulieu, 1997; Belkin et al., 2003; Ruthven, 2003). On the other hand, there are some indications that the performance can be improved when a greater number of candidate terms are presented (Belkin et al., 2003; Jones et al., 1995). The next section discusses the limitations of existing approaches to IQE.

## 3.4 Limitations

Several characteristics and limitations of the existing approaches to IQE can be identified from the studies reviewed so far. First of all, the vast majority of studies in IQE have been based on a RF approach where users were asked to make a relevance judgement of retrieved documents before a system suggested a set of candidate expansion terms. RF techniques have been extensively investigated in both real user or simulated user studies. While Belkin et al. (2003) show that an LCA based approach could be as effective as a RF approach, the evaluation of alternative QE approaches appears to be limited.

As mentioned above, a limitation of traditional RF techniques is that users have to explicitly indicate the relevance of some of retrieved documents to get candidate terms from a system. This method forces the users to engage in additional actions, thus, it can increase the users' cognitive load during the search. In order to overcome this problem, techniques such as pseudo RF (which assumes that the top N retrieved documents are relevant) and implicit feedback (which attempts to infer relevance information through interaction) have been studied (Kelly and Teevan, 2003). Also, an optimal size of candidate expansion terms can vary across topics, but existing techniques tend to offer a fixed number of terms to the users.

Another aspect to be addressed is the source of expansion terms. Several studies have exploited knowledge structures such as thesauri. A thesaurus and its hierarchical structure can be useful for providing the semantic relationships between terms. However, most studies have tended to focus on what semantic relations can improve retrieval effectiveness. The general findings suggest, though, it largely depends on the topics. In addition, it is expensive to build and maintain a knowledge structure manually, and they are not available in every domain. A limited coverage of vocabulary can cause a difficulty in generating effective candidate expansion terms (Jones et al., 1995; Voorhees, 1994).

The limitations are closely related to the presentation of expansion terms. Regardless of the techniques or sources being used, most existing approaches to IQE can be seen as a ranking approach where terms are ranked and a relatively small set of the top scored terms are presented to the users. A form of lists is typically used for the presentation. As we have discussed above, the main limitation of lists is the lack of contextual information that should support a user's selection of terms for query reformulation. Ruthven (2003) also argues that "the lack of connection between expansion terms and documents used to provide those terms indicates that searchers may need more support in how to use query expansion as a general interactive technique".

More specifically, the following aspects of candidate terms appear to be missing from the presentation of existing ranking-based approaches:

- the relationship between query terms and candidate terms
- the relationship between candidate terms
- the relationship between candidate terms and documents from which they are derived

Finally, as Efthimiadis (1996) pointed out, more investigations are required to understand users' searching behaviour. The lack of users studies with different approaches is the general problem and limitation of IQE research to date.

## 3.5 Summary

This chapter has provided a literature review of studies related to query reformulation. We surveyed a range of methods and sources that have been used for QE. The major sources of expansion terms were categorised into user feedback, retrieved documents, global collection, and manually constructed knowledge structures. As for the QE methods, approaches such as relevance feedback, term clustering, and LCA have been discussed as well as the techniques based on a thesaurus. It was pointed out that the vast majority of QE studies have been based on relevance feedback, and limited work has been carried out with different approaches.

We also looked at the HCI aspects of query reformulation, where we have discussed studies based on both real users and simulated users. The simulations were found to be useful to understand the potential effectiveness of IQE. However, the comparison with the AQE performance suggests that a simple conversion of an AQE into an interactive fashion does not always allow users to select useful terms. This observation was also supported by several studies which investigated the information seeking behaviour of users in the context of QE. It was also pointed out that the presentation of candidate terms has received little attention so far.

Lastly, we discussed the limitations of existing approaches to query reformulation, and suggested several factors that motivated our approach to interactive query reformulation. One of the main criticisms emphasised in this section was the lack of structure and context in the selection and presentation of expansion terms. The problem also appears to limit the range of potentially useful terms which can be offered to users.

To address some of the problems discussed in this chapter, the next chapter investigates studies related to automatic generation of an overview of retrieved documents.

# Chapter 4

# Deriving a hierarchical structure of concepts from retrieved documents

## 4.1 Introduction

The previous chapter has reviewed existing QE techniques and discussed how automatic QE methods can be applied to interactive query reformulation. It was pointed out that there was room for improving the interaction between an IR system and users in terms of query reformulation. Particularly, the lack of contextual information of suggested terms offered to support a user's term selection was highlighted as an area to address. It was also pointed out that a simple structure such as lists has dominated the presentation of suggested terms in existing IQE approaches, and that limited work has been carried out to investigate the effectiveness of other presentations. We argued that the frequent use of lists could be partly due to the underlying QE techniques which were often designed to rank candidate terms. However, we speculated that a ranked list might not be the best presentation to offer sufficient contextual information of suggested terms. Therefore, a new approach should be investigated to derive a better structure of suggested terms to overcome the limitation of existing IQE approaches.

In this thesis, we investigate a concept-based approach to support a user's query reformulation and searching using a hierarchical structure of concepts, that is derived from a set of retrieved documents. Like a topic directory available on the web

(e.g., Yahoo![1] and Open Directory Project[2]), we are interested in exploiting a hierarchical structure where a more general concept is located at a higher level of the hierarchy, and related but more specific concepts are located at a lower level of the structure. However, we are also interested in automatic generation of the hierarchy based on a set of retrieved documents. The motivations for the use of such a hierarchy to support a user's query reformulation and searching are as follows.

Firstly, hierarchies enable us to represent a relationship between terms using the structure. Given that a relationship between query terms and suggested terms was not well represented in existing IQE approaches, this can be an advantage for increasing a user's awareness of the relationship between terms. Secondly, hierarchies allow us to free from a fixed number of terms presented to a user. Since the number of useful terms can vary across topics and there can be many terms that are potentially relevant to a topic, a hierarchy appears to be more suitable for presenting a wide range of terms in a structured way, compared to a long list. Thirdly, hierarchies can have a structure for navigating a user from a general concept to related but more specific concept (Hearst, 1999). While a user might find some of suggested terms unfamiliar, a hierarchy can help a user infer the meaning of unfamiliar terms based on a parent term in the hierarchy or other terms in the same node. The hierarchical navigation is also an advantage of using a hierarchical structure compared to a list. Lastly, by deriving a hierarchy from a set of retrieved documents, we can offer a means of getting an overview of retrieved documents to a user via the browsing of terms in a hierarchy.

As can be seen, there appears to be several advantages of using a hierarchy to support a user's query reformulation and searching, compared to a simple structure such as lists. An empirical study also suggests that a hierarchy can be more efficient for selecting terms than a list when a wide range of terms is presented to a user (Joho et al., 2002). In addition, our approach of using retrieved documents to derive a hierarchy appears to be supported by the finding of Efthimiadis (2000), which suggests that a user can make use of terms whose relationship with a query term is not necessarily defined by a thesaurus. Therefore, our approach is likely to be

[1] `http://www.yahoo.com` [Accessed: 6/1/07].

[2] `http://www.dmoz.org` [Accessed: 6/1/07].

domain-independent and can be applied to a wide range of queries that might be requested to a large corpus.

The rest of this chapter reviews studies related to automatic generation of a hierarchical structure, which we refer to a *concept hierarchy*. We discuss studies from three perspectives: extracting concepts, identifying a relation between concepts, and presenting a hierarchy. Note that a *concept* is used to refer to a word or phrase.

## 4.2 Extracting concepts

### 4.2.1 Definition of concepts

Extraction of concepts from texts involves implicitly or explicitly a task of defining a concept. If a "bag of words" approach is used, any word, which can be seen as a text string separated by non-word characters (Grefenstette and Tapanainen, 1994), would be a concept with a possible exception of stopwords. While a bag of words approach could introduce a large amount of irrelevant concepts, it could be computationally efficient.

An alternative approach is to define a concept as a syntactic construction, or sequence of particular parts-of-speech. For example, noun phrases are frequently used as a useful unit of concepts in text-based applications (Anick and Tipirneni, 1999; Grefenstette, 1992; Wacholder et al., 2001; Xu and Croft, 1996). Anick and Vaithyanathan (1997) provide a summary of advantages of using noun phrases as opposed to other syntactic constructions, as follows:

- Noun compounds are widely used for describing concepts succinctly;
- because they are contiguous, they are relatively easy to detect and extract from text corpora;
- unlike many phrasal constructions, which reflect transient relationships among objects, noun compounding is generally applied to express tighter, more long-lived relationships between concepts, thereby contributing less noise;
- most proper nouns are subsumed under this definition of noun compounds; and finally,

- nouns and noun compounds form the bulk of the terms that show up in actual queries.

Extraction of noun phrases usually involves a tokenisation of texts, sentence identification, parts-of-speech tagging, and finally extraction of defined sets of parts-of-speech. We have seen how a tokenisation worked in Section 2.2. Identification of sentence boundary helps a system to avoid extracting words at the end and beginning of sentences as a phrase (e.g. "... was beautiful. Castles are ..."). A simple technique for splitting texts into sentences is to use punctuation marks such as ".", "!", and "?", along with a list of abbreviations such as "Dr.", "Sep.", or "Plc." to reduce error of splitting (See Palmer and Hearst (1997) for more details).

### 4.2.2 Parts-of-speech tagging

The next stage involves an identification of parts-of-speech (POS) for each of the tokens in a sentence. One of the simple approaches to POS tagging is based on a lexicon, which is a list of words with a set of potential POS tags found in a training corpus. Lexicons manually generated from Brown Corpus, Wall Street Journal, and British National Corpus are popular in English. Unknown words that do not appear in a lexicon can be estimated by grammatical contexts or can be processed as a proper noun. Brill (1992) shows that tagging with the most frequent parts-of-speech found in a training corpus can achieve approximately 92% of accuracy, while a rule-based technique improves it by 3%. Other techniques appear to achieve a similar region of accuracy. Once POS tagging is completed, a set of pre-defined sequence of speech (e.g., Noun + Noun, Adjective + Noun) can be extracted as noun phrases.

## 4.3 Identifying a relation between concepts

The core technology behind the generation of concept hierarchies is to determine a relationship between concepts. While there are a number of techniques to estimate relatedness (or similarity) between concepts, the construction of concept hierarchies requires a further step of identifying a type of relationship between concepts. This

section reviews several techniques to determine a hierarchical relationship of concepts from texts. The techniques are broadly divided into a statistical approach and lexical approach.

### 4.3.1 Statistical approaches

A statistical approach to derive a hierarchical relation between concepts often makes use of word frequency, document frequency (DF), and co-occurrence information. Applications based on term clustering discussed in Section 3.2 are one such example. The basic idea of term clustering is to determine the similarity between terms by measuring the number of document in which they co-occur. This derives a list of associated terms, which can also be seen as synonyms of a term.

By linking strongly associated terms, one can generate a graph as illustrated in the left of Figure 4.1. Each node is a single concept. One way to derive a hierarchical structure from the graph is to order the concepts based on a variant of document frequency (shown in the right side of Figure 4.1). The second figure illustrates that the graph is roughly divided into a hierarchy which comprises of three levels. The levels are based on a variant of document frequency (DF) of concepts in a collection. This approach has been used by Niwa et al. (1997) to offer a visual link between concepts and search results, and by Nanas et al. (2003) to offer a set of documents that are different from an existing personal collection in information filtering.

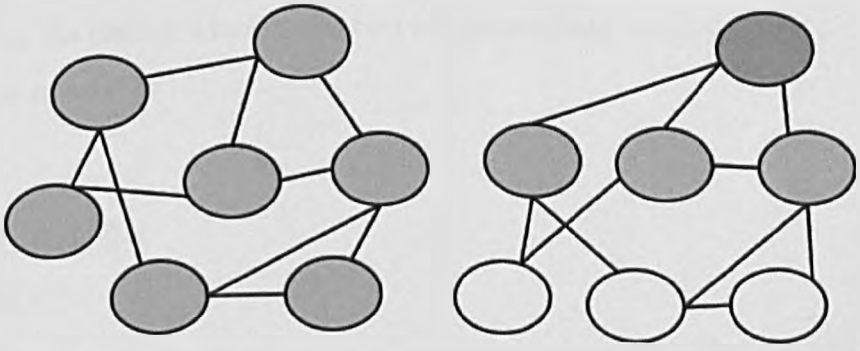


**Figure 4.1:** *A topic graph (Left) and its hierarchical order based on document frequency (Right). A higher document frequency term is represented in a darker colour.*

Sanderson and Croft (1999) developed an alternative technique that attempts to determine a *subsumption relation* between terms. The following example illustrates the difference between a simplified term clustering and Sanderson and Croft's ap-

proach. Given that two concepts $t_1$ and $t_2$ be formalised as:

$$t_1 = \langle d_1, d_2, \ldots, d_n \rangle \tag{4.1}$$

$$t_2 = \langle d_1, d_2, \ldots, d_m \rangle \tag{4.2}$$

where $n$ is the number of documents indexed for $t_1$, and $m$ is for $t_2$, term clustering focuses on a set of documents commonly indexed by two terms to measure a similarity between $t_1$ and $t_2$. The more documents two terms share the more similar they are assumed to be.

A subsumption hierarchy, on the other hand, examines if $t_1$'s set of the documents subsumes $t_2$'s. More specifically, $t_1$ is said to subsume $t_2$ when the following two conditions hold:

$$p(t_1|t_2) = 1 \tag{4.3}$$

$$p(t_2|t_1) < 1 \tag{4.4}$$

The assumption is that $t_1$ is likely to be more general than $t_2$ because, firstly, the former occurs in a set of documents more frequently than the latter, and secondly, the former subsumes a large part of latter's document set. Also, they are likely to be related since they co-occur frequently in a set of documents. The authors, however, found that the chance where these two conditions hold was relatively rare; hence, they were relaxed to:

$$p(t_1|t_2) \geq 0.8 \tag{4.5}$$

$$p(t_2|t_1) < 1 \tag{4.6}$$

where 0.8 was empirically determined by Sanderson and Croft (1999) after some experimentation of subsumption term pairs. The original and relaxed conditions are illustrated in Figure 4.2.

Sanderson and Croft (1999) comment that a hierarchy seems better when the co-occurrence is measured based on paragraphs as opposed to an entire document.

An association analysis of randomly selected pairs shows that nearly half of the cases were found to have some sort of relationship. Of the relevant pairs, 49% were judged as *aspect of*, 23% as *type of*, 8% as *same*, 1% as *opposite*, and finally, 19% as *don't know*. Given that *aspect of* and *type of* can be an appropriate parent-child relation, 72% of relevant pairs were found to form a hierarchical relation.

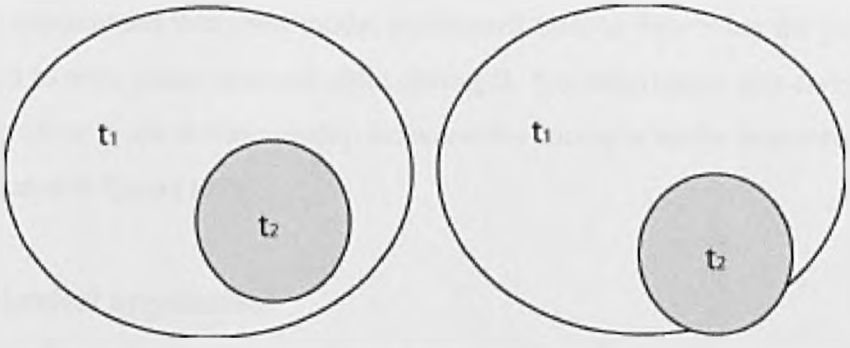



**Figure 4.2:** *Illustration of Subsumption relationship (Left: Original condition, Right: Relaxed condition.*

The main difference between the topic graph and subsumption hierarchy is that association of terms in the former approach is designed to be symmetric while the latter is designed to be asymmetric. The asymmetry provides a context of concepts. For example, if a pair of concepts, *UK* and *economy*, is defined as a parent and child in a subsumption hierarchy, it is likely that the concept *economy* is of the country *UK* since the majority of *economy*'s documents are a subset of *UK*'s. In other words, the latter model is aimed to determine a parent-child relationship more clearly than the former.

The common feature among the statistical approaches is that they use a variant of DF to determine the specificity of concepts. For example, both Sanderson and Croft (1999) and Nanas et al. (2003) refer to the paper by Forsyth and Rada (1986) who proposed to assume that DF should be diagnostic of term specificity. Sparck Jones also gives some accounts on the relationship between term specificity and DF in her paper which introduces the notion of IDF Sparck Jones (1972). Since this is one of the core factors in the statistical approaches of concept hierarchies, we will investigate this aspect in depth in a later chapter of this thesis.

In addition to the notion of a parent and child, Glover et al. (2002) attempted to identify the notion of *self* (i.e., synonyms) by using two types of IDF scores: one was based on a global collection (global IDF, or GIDF) and another was based on a

set of retrieved documents (local IDF, or LIDF). In their model, a self was defined as a relatively low GIDF and very high LIDF, parent was defined as a middle range of GIDF and LIDF, and finally, child was defined as a very low GIDF and a middle range of LIDF. The actual cutoff values were estimated by a distribution of concepts extracted from the top levels of Open Directory Project's topic directory. The authors commented that their model performed well to determine the parents but appeared to miss many relevant child concepts. No experiment was carried out to examine other types of relationship between the concepts as the one conducted by Sanderson and Croft (1999).

### 4.3.2 Lexical approaches

Another body of research on deriving a hierarchical concept structure is based on lexical analysis. For example, Anick and Tipirneni (1999) looked at nouns that were frequently used in noun phrases in a corpus. Their method is based on the lexical dispersion hypothesis, which is, "the number of different compounds that a word appears in within a given document set can be used as a diagnostic for identifying key concepts of that document set". Some example terms identified by their method were as follows.

- Source: Financial Times
  - Market: stock market, bond market, gilt market, market share
  - Group: insurance group, industrial group, Pegasus Group
  - Company: public company, car company, company formation
- Source: A Jazz music database
  - Music: world music, sheet music, music director
  - Jazz: Dixieland jazz, jazz ensemble, Antibes Jazz Festival
  - Band: swing band, band leader, Robert Cray Band

Those noun phrases were then categorised under the key terms (i.e. market, group, company). When it was required to rank the key terms, for example, for query expansion, they used frequency of occurrence of the key terms in the compounds. However, since one noun phrase might appear in a single long document

many times but not in other documents, a threshold was used to stop the counting of occurrence in the same document. In their study, the threshold was set to five.

Grefenstette (1997) proposed to present noun and verb phrases based on contexts. Nine proposed contexts were based on syntactic roles of words that modified another word in phrases. Basic NLP techniques such as tokenisation, morphological analysis, and part-of-speech tagging were applied to texts in advance. Examples of their contexts using a concept *research* are shown in Table 4.1.

**Table 4.1:** *Grefenstette's nine contexts of phrases*

| Phrase | Context | Examples |
|---|---|---|
| Noun | Types of research | Market research, recent research, scientific research, extensive research, research institute |
| | Research things | Research project, research centre, research team, research institute |
| | Named research | American research, NHS development research, EC research, Xerox research, research institute |
| | Names involving research | Medical Research Council, Cancer Research Campaign, Defence Research Agency |
| Verb | Things one can research | Research book, research project, research report, research design |
| | Who can research | Student research, team research, sociologist research, computer research |
| | How can one research | Research carefully, research extensively, research recently, research relatively |
| | Things that one does to research | Conduct research, support research, publish research, complete research |
| | What can research do | Research show, research indicate, research provide, research concentrate |

As can be seen, the contexts were determined mainly depending on parts-of-speech of the modifiers. For example, *type of research* grouped noun phrases consisting of `adjective` followed by research, and *research thing* grouped research followed by `noun`. *Named research* and *Names involving research* were based on capital letters of the first letter of words. Similar patterns were applied to verb phrases. Note that the same phrase can be found in more than one context in his model. For example, the concept *research project* was found in the context of *Research things*

as well as *Things one can research.* Grefenstette explained that the contextual labels were assumed to be easier for average search engine users to understand than categorisation based on syntactic roles alone. However, no empirical study was carried out to investigate the validity of such an assumption.

Woods (1997) organised concepts by identifying relations between head nouns and modifiers in phrases. His method used various NLP techniques and resource for the identification of relations, although the basic idea was relatively straightforward. For example, a noun phrase *automobile cleaning* can be seen as a compound of a modifier (i.e., automobile) and head noun (i.e., cleaning), like Grefenstette's method. Woods categorised the phrase as a kind of *cleaning* since it was modified by something (i.e., automobile). *Car washing*, on the other hand, was organised under *automobile cleaning* since a car was a kind of automobile. An example of his conceptual taxonomy is shown in Figure 4.3.

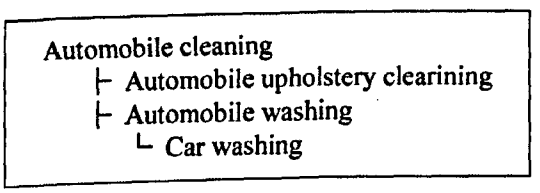



**Figure 4.3:** *An example of Woods' conceptual taxonomy.*

As can be seen, Woods' approach was mainly based on IS-A relations between head nouns and modifiers (i.e., Washing is a kind of cleaning; A car is an automobile, etc.). The process was repeated for phrases that had more than one modifiers. For example, *industrial steam cleaning* was organised as a child concept of *steam cleaning.*

A distinct technique to derive a hierarchical relationship is Lexicon-Syntactic Pattern Extraction (LSPE). The LSPE was originally proposed by Hearst (1992) to populate additional parent-child relations to WordNet (Miller, 1995). An advantage of LSPE was that it did not require a knowledge-based algorithm or tool to identify hierarchical relationships. Instead, simple text fragments were used. An example text fragment was *such as*, which may be found in a sentence like the following:

> "... used several search engines such as Google, AltaVista, and Goo in order to compare the performance of ..."

Even if one did not know *Goo*, s/he was likely to infer that *Goo* was one of the search engines[3]. Similarly, several text fragments such as *and other* and *or other* were used as the textual patterns in LSPE. From the linguistic point of view, Hearst's method can be seen as extraction of super-ordinate concepts (referred as *hyponym* in the original paper). Radev (1998), on the other hand, attempted to extract appositive concepts. The example may be found in the following sentence.

> "Sun Microsystems, the leading workstation manufacturer, announced a new series of ..."

The phrase *The leading workstation manufacturer* was an appositive phrase of *Sun Microsystems*, and provided a description of the company. In other words, "appositives are used in English to further specify the meaning of the noun they follow", and also "anything can be described by an appositive, and the method of deriving the descriptive information is the same in every case" (p. 447) (Coates-Stephens, 1993). The linguistic pattern also allows us to locate a parent-child description in texts. Since LSPE and Radev's pattern together did not require intensive NLP processes unlike the other lexical approaches discussed before, it appears to be a promising alternative approach to build a concept hierarchy in a domain-independent application. We expand this approach and evaluate the performance in a wider context at a later chapter of this thesis.

## 4.4 Presenting the concept structure

We have discussed several techniques which can be used to derive a concept hierarchy from texts. We have seen that both statistical and lexical attributes can be exploited to determine a hierarchical relationship between concepts. However, these techniques are not necessarily designed for users to browse the derived structure in an interactive fashion. Therefore, this section explores different types of visualisation which can be used to present a concept hierarchy to users, in the context of supporting their query reformulation and searching.

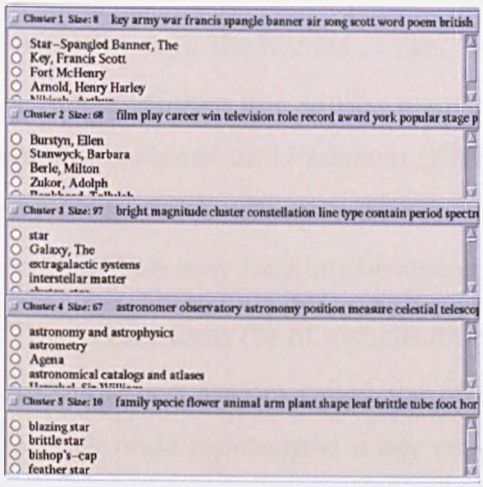
[3]Goo (`http://www.goo.ne.jp` [Accessed: 6/1/07]) is a search engine available in Japan.

In the introduction of this chapter, we discussed potential advantages of our concept hierarchy approach compared to a ranked list which has been typically used to present candidate terms in existing IQE approaches. The advantages and objectives of our approach can be summarised as follows.

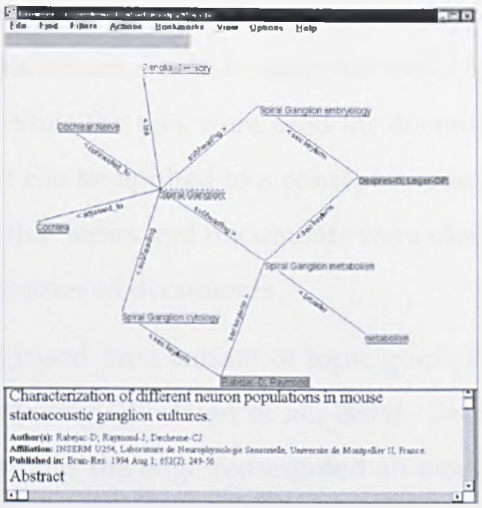
- Representation of a relationship between terms. Hierarchies enable us to represent a relationship between query terms and suggested terms, and between suggested terms. Represented relationship can be seen as additional contextual information of suggested terms offered in our approach.
- Representation of term specificity. Hierarchies allow us to represent a level of term specificity of suggested terms by organising a more general term at a higher level and related but more specific term at a lower level of the structure. Therefore, our approach is designed to navigate a user from general to specific concepts.
- Diversity of potentially relevant terms. By using such a structure, we speculate that hierarchies can be more suitable to present a wide range of terms potentially relevant to a topic than a long list. Therefore, it might be useful for increasing a user's search vocabulary employed to complete a search task.
- Overview of retrieved documents. A range of terms populated in a hierarchy can offer an overview of retrieved documents to a user. Therefore, an effective way of browsing suggested terms is important in our approach.

Presentation of a concept hierarchy is, therefore, an important aspect of our investigation. The rest of this section explores some of existing methods developed to visualise a hierarchical structure of concepts. Our aim of this section is to analyse the characteristics of visualisation methods and assess their possibility of facilitating the objectives of our approach summarised above.

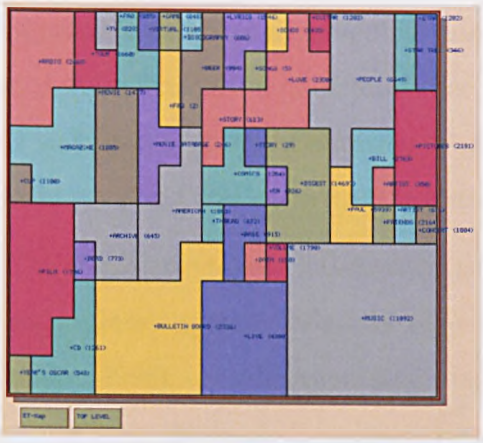
Figure 4.4 shows six forms of visualised structures which can be found in the literature. The first form (a) of visualisation is based on a text tile. This form was proposed by Hearst and Pederson (1996) for the visualisation of a hierarchical document clustering technique called the *Scatter/Gather* (Cutting et al., 1992). The tiles were vertically arranged and ordered by the size of clusters. In each tile, the title of



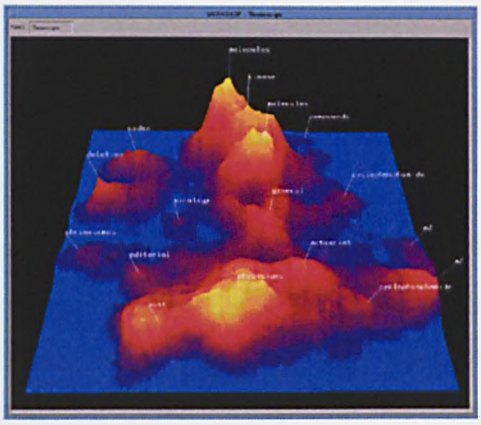

(a)



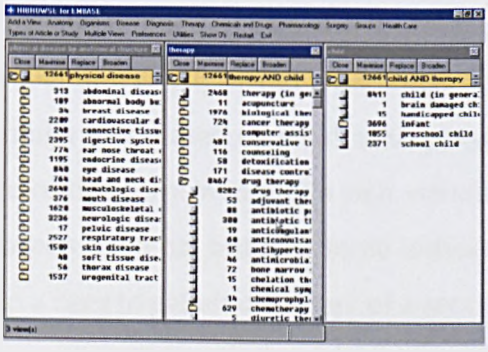

(b)



(c)



(d)



(e)


(f)

**Figure 4.4:** *Visualisations of concept structures (a: Tile structure (taken from Hearst and Pederson (1996), b: Topic graph (taken from Wiesman et al. (2004)), c: 2D Map structure (taken from Chen et al. (1998)), d: 3D Map structure (taken from Wise et al. (1995)), e: Tree structure (taken from Pollitt (1997)), f: Cascading menu structure (taken from Sanderson and Croft (1999)))*

documents in the cluster was listed. Representative terms extracted from each cluster were shown at the header of tiles. When a user selected one of the tiles, child clusters were shown in a similar manner. While the tiles were used for document clustering in Hearst and Pederson (1996), it can be applied to a concept hierarchy. An advantage of the tile visualisation was that terms and documents were closely associated. This may facilitate browsing of retrieved documents.

The second form (b) of visualisation is based on a variant of topic graph and the example is a system called the *Metabrowsing* (Wiesman et al., 2004). In this form, each node represented a key concept and the edge represented an association between concepts. The advantage of this form was that an association between concepts was well represented. In the example based on a manually constructed bibliographic database in a medical domain, the concepts such as the name of disease, organs, and authors of published papers were linked along with a relation label such as a *broader, subordinate,* or *connected to.* A set of documents in which a concept occurred was shown at the bottom of the interface.

The third form (c) of visualisation is based on a two dimensional (2D) map and the example is a system called the *self-organising map* (Kohonen, 1995; Deboeck and Kohonen, 1998). A self-organising map (SOM) provided "a topology-preserving mapping from the high-dimensional space to map units", hence, one can create a SOM based on a term-document multi-dimensional space (Chen et al., 1998). The example map was based on a set of web pages registered in the Entertainment category in Yahoo!. The size and location of concept regions were determined by the number of web pages and related concepts. When a region of the map was selected, a similar map based on child concepts was shown in a similar manner. The advantage of this form was that a rich visualisation represented not only a relationship between concepts but also some indication of term specificity. Thus, it appears to help a user to get an overview of a set of documents.

The fourth form (d) of visualisation is based on a three-dimensional (3D) map and the example is a system called the *ThemeScape* (Wise et al., 1995). Similar to SOM, ThemeScape was based on a term-document multi-dimensional space, and related concepts were arranged in a shorter distance than others were. The difference from the SOM's two-dimensional visualisation was that ThemeScape repre-

sented significance of concepts as a height. Furthermore, features such as valleys, peaks, cliffs, and ranges were designed to represent detailed inter-relationships among the concepts and documents. The example map was based on 20,000 documents from CNN news. The advantage of this form of visualisation was that a rich visualisation could represent complex relationships between concepts like SOM.

The fifth form (e) of visualisation is based on a hierarchical tree and the example is a system called the *HIBROWSE* (Pollitt, 1997). HIBROSE was designed to facilitate browsing of search results where the documents were associated with multi-faceted concepts from a thesaurus. In the example, the tree structures were generated based on a manually constructed hierarchical classification scheme, and displayed for each of the facets defined by users. In each tree, concepts were ordered alphabetically along with the number of associated documents. The presence of child concepts was indicated by the folder icon in the interface. The advantage of this visualisation was the simplicity of structure. A hierarchical relationship between concepts was also well represented.

The final form (f) of visualisation is based on a cascading menu and the example is a system called the *subsumption hierarchy* (Sanderson and Croft, 1999). This visualisation is similar to the tree structure. A hierarchical relationship was represented by sub-menus, thus, concepts that were more specific could be browsed with a clear link to related but more general concepts. The menu system also appears to be familiar to users since it is used in the graphical user interface of various applications.

As can be seen, six forms of visualisation appeared to share some advantages. For example, all forms were likely to be able to present a wide range of concepts derived from a set of documents. Also, they appeared to help a user get an overview of documents. As Hearst (1999) suggests, such an overview appears to "help users get started, directing them into general neighbourhoods, after which they can navigate using more detailed descriptions (p.268)". However, there were also different advantages among them. In terms of relationship between concepts, the Metabrowsing, SOM, and ThemeScape appeared to be superior to other forms of visualisation in representing a wide range of relationships. The SOM and ThemeScape also had an advantage of graphically representing the specificity and/or significance of con-

cepts in document space. However, the graphically rich representation can be too complex for a user to fully take advantage of their utility. Empirical studies also suggest that a complex structure is not always effective to support an IR system's user (Chen et al., 1998; Sebrechts et al., 1999). Another disadvantage of the rich visualisation is that it takes up large amounts of screen space, which is a common problem in the interface design with a hierarchical structure (Hearst, 1999).

The tree and menu forms, on the other hand, appeared to have an advantage of presenting a wide range of potentially relevant concepts with concise visualisation. While the structure of the two forms was simpler than other forms, they seemed to be sufficient to achieve the objectives of our approach. Both forms provided an intuitive way of representing a hierarchical relationship between concepts. The level of depth in the structure can also represent specificity of concepts in document space. In addition, since their visualisation is simple, we speculate that users should be able to browse the concepts to get an overview of retrieved documents. Therefore, the forms appeared to be promising for our application.

The interface of the Scatter/Gather, Metabrowsing, and SOM also suggests an approach to facilitating an association between a concept and documents using a visualised hierarchical structure. The interface was designed to show a group of documents in which a concept was associated. A similar interaction can be achieved in our approach. In other words, when a concept was selected in hierarchies, a set of retrieved documents in which the concept occurred could be presented to a user. This can give users a control on how retrieved documents were grouped and explored during a search task. Therefore, a concept hierarchy can be used to support a user's browsing of retrieved documents. We speculate that the grouping of retrieved documents using a concept hierarchy can also help a user clarify the relationship between concepts in retrieved documents.

As can be seen, the visualisation of concept hierarchies discussed in this section suggests that there can be multiple aspect of information searching process that can be facilitated by our approach. One is a user's search vocabulary employed to complete a search task. By providing further information on the relationship between query terms and suggested terms, our approach aims to facilitate a user's term selection from a wide range of potentially relevant terms. The hierarchical

presentation of candidate terms is also used to navigate users from a general concept to related but more specific concept found in retrieved documents. Another aspect is a user's browsing of retrieved documents. By grouping retrieved documents using selected hierarchy terms, our approach aims to give users a control on how documents are organised and explored to complete a search task.

In Chapter 8, we develop a search interface where a visualised hierarchy is incorporated into an interface of an IR system. The findings from this section are considered in the design of our search interface and evaluation of our approach to supporting a user's query reformulation and searching.

## 4.5 Related work

While this chapter has focused on work which was potentially useful for supporting query reformulation and searching in IR, deriving a structure from texts and its visualisation have a large body of research and related studies in other fields. This section discusses one of such studies.

### 4.5.1 Phrase browsing

Phrase browsing applications have recently been investigated actively in the context of digital libraries. In the workshop held in the first ACM-IEEE Joint Conference on Digital Libraries, the organisers of the workshop described the role of the phrase browsing applications in the digital libraries as follows:

> "Phrase browsing applications provide information seekers with access to text content via structured lists of index terms. The index terms, which may be identified by a variety of techniques, are phrases that have been automatically extracted from full text documents. Browsing applications support interactive navigation of index terms and provide direct access to the original documents via the index terms. Terms are presented to users in ways that allow them to either 'drill down' from a shorter, more general terms to longer, more specific ones or to navigate

from one term to other related ones via graphical interfaces (Wacholder and Nevill-Manning, 2001)."

As can be seen, there are several aspects that are common with our objectives. For instance, we both aim to derive a structure from texts in a presentable form to users to assist them in getting a better idea of the contents. A hierarchical navigation was considered as a part of graphical interfaces of the phrase browsing. The main difference from our study was that the phrase browsing applications were primarily designed for an entire document collection as opposed to a set of retrieved documents. Therefore, exhaustiveness of extracting index terms from texts appears to be a prominent issue for them. The following is an example of phrase browsing applications.

Paynter and Witten (2001) developed an interface of the phrase browsing which combined vocabularies manually defined by a thesaurus and phrases extracted from a document collection as index terms. Their interface provided a lookup function for the combined list of index terms in addition to the standard searching function. When the lookup function was called, the interface firstly showed matched terms in the thesaurus, along with the broader terms, narrower terms, or related terms defined by the thesaurus. Each record also showed document frequency and word frequency of the terms in a document collection. The interface further showed a list of extracted phrases in which the terms occurred. When a user selected one of the phrases from the list, an expanded list was shown if there was any phrase that contained the selected phrase. The title of the documents in which the selected phrase occurs was also displayed below.

## 4.6 Summary

This chapter has reviewed techniques to derive a hierarchical structure of related concepts from retrieved documents. We argued that hierarchies had several advantages to overcome the limitation of existing IQE approaches. In particular, representation of relationship between concepts, representation of term specificity, ability to present a range of potentially relevant concepts, and ability to offer an overview of

retrieved documents were highlighted as a means of providing additional contexts of suggested terms using hierarchies.

The rest of the chapter was divided into three conceptual stages of automatic generation and use of concept hierarchies. In the first stage, we have discussed extraction of concepts. It was pointed out that extraction of concepts involved the definition of concepts. Several types of potential definitions were introduced and their characteristics were discussed. Particularly, noun phrases were argued as a useful unit of concepts, and the basic procedure to extract noun phrases was shown.

The association and relationship between concepts were discussed in the second stage. Both statistical and lexical approaches to derive a hierarchical structure from texts have been investigated. In statistical approaches, techniques based on co-occurrence information were examined and the subsumption hierarchy was highlighted as a distinct way of using co-occurrence information. In lexical approaches, techniques based on phrase analysis were investigated where different levels of natural language processing and use of lexical resources were involved.

The last stage involved the visualisation of concept structures. The characteristics of six forms of visualisation were discussed in the context of supporting a user's query reformulation and searching. It was suggested that a relatively simple visualisation method can be sufficient to achieve the objectives of concept hierarchies in our approach, although rich visualisation might be able to represent more complex relationships between concepts. It was also pointed out that the effectiveness of concept hierarchies to facilitate browsing of retrieved documents was worth investigating. These findings will be considered in the design of our interface and evaluation of our approach presented in Chapter 8.

The review of automatic generation of concept hierarchies highlighted two major issues which required further investigation. The first area was the relationship between frequency of occurrence of concepts and their specificity. It was pointed out that statistical approaches to concept hierarchies have often assumed an implicit correlation between them. However, limited work has been carried out to investigate their relationship using a large corpus. The second area was an application of Lexicon-Syntactic Pattern Extraction (LSPE) in extraction of parent-child descriptions. While LSPE appeared to be a promising technique to find a hierarchi-

cal relationship from texts, it has only been used in a limited application. These two issues are investigated and evaluated in Chapter 6 and 7, followed by a summative evaluation of our approach presented in Chapter 8. The next section discusses an overview of experiments presented in the three chapters.

# Chapter 5

# Overview of experiments

This thesis aims to present and evaluate a concept hierarchy approach to a user's query reformulation and searching in interactive IR (IIR). Compared to the existing ranking-based query expansion techniques, our approach is designed to offer a wider range of potentially relevant terms to the user. Our approach also aims to offer the contexts of suggested terms by using a hierarchical structure in the presentation of terms. In addition, we hope that our approach can help the user get an overview of retrieved documents by browsing the concept hierarchies.

This thesis particularly looks at the automatic generation of concept hierarchies based on a set of retrieved documents, to achieve a domain-independent application that can be used in an IR system with a heterogeneous document collection. The previous chapter discussed the methods of the automatic generation of concept hierarchies, and suggested the venues for further investigation. This investigation related to the generation of concept hierarchies is carried out by corpus-based analysis. Furthermore, a summative evaluation of our approach is carried out by a task-based study with real users. This chapter provides an overview of the experiments, which will be presented in the next three chapters.

**Experiment I (Chapter 6):** The goal of Experiment I is to gain further insight into the relationship between the document frequency and term specificity. The statistical methods of generating a concept hierarchy often appear to assume that a term's frequency of occurrence can be used to indicate its semantic specificity (e.g., Niwa

et al., 1997; Sanderson and Croft, 1999; Nanas et al., 2003). However, limited work has been carried out to examine their relationship using multiple large scale corpora. In Chapter 6, varied size of corpora from the TREC collections and a search engine are used to estimate the document frequency of terms in a thesaurus. As a result, we present an empirical study investigating the relationship between the document frequency and term specificity, using larger corpora and wider range of terms than existing work such as Weinberg and Cunningham (1985) and Caraballo and Charniak (1999).

**Experiment II (Chapter 7):** The goal of Experiment II is to evaluate a lexical approach to find parent-child descriptions from texts. Our intention is to derive a concept hierarchy using the parent-child descriptions, which can be compared to a statistical approach in Experiment III. In Chapter 7, we develop an extraction system by expanding the Lexicon-Syntactic Pattern Extraction (LSPE) method originally proposed by Hearst (1992). The LSPE method does not require an intensive NLP process or large knowledge-base resource; thus, it can be used in a domain-independent application. The system is designed to enhance the lexicon-syntactic patterns with other evidence derived from a corpus. The experiment investigates the effectiveness of the individual lexicon-syntactic patterns and the effect of corpus size. It also illustrates how a concept hierarchy can be generated from the descriptions.

**Experiment III (Chapter 8):** The goal of Experiment III is to evaluate the effectiveness of our approach to a user's query reformulation and searching. As opposed to the corpus-based analysis in the previous two experiments, Experiment III is a task-based user study. In Chapter 8, we develop a search interface which presents a concept hierarchy along with search results. The interface allows users to browse the concept hierarchy to reformulate an existing query. It also enables users to access a subset of retrieved documents based on the terms in the hierarchy. Chapter 8 describes the detail of our implementation of the interface based on two types of concept hierarchies: a statistical technique and lexical technique. The two types of hierarchy are compared to a baseline interface which has no hierarchy. The user study is carried out using one of the TREC test collections specifically designed for

the evaluation of IIR systems.

As can be seen, the three experiments presented in the following chapters have a different research goal and experimental design. However, the TREC test collections form the basis of our investigation across the experiments.

# Chapter 6

# Experiment I: Document frequency and term specificity

## 6.1 Introduction

The previous chapter gave an overview of three experiments which will be presented in the following chapters. This chapter presents the first experiment which investigates the statistical aspects of concepts regarding the automatic construction of concept hierarchies. In particular, we investigate the relationship between document frequency and term specificity.

Topic hierarchies have long interested researchers in information retrieval (IR), computational linguistic (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 of research, 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 (Anick and Tipirneni, 1999; Grefenstette, 1992).

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 might not correlate strongly: Sparck Jones (1972), in her paper introducing IDF, implied that the two were not the same thing. However, the focus of her paper was not on specificity but on evaluating the retrieval effectiveness of a term weighting based on IDF, thus, 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 type4) was limited.

The work in this chapter 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 chapter 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 chapter firstly reviews the related work. Next, the design and

results of the experiment examining specificity in document collections is outlined, followed by the experiment focusing on sets of retrieved documents. The implications of the results for the development of topical hierarchies are then discussed before the chapter concludes.

## 6.2 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, as 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 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 might 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. Sparck Jones commented that this type of specificity was not necessarily the same as a semantic perspective but it was useful for retrieval systems. In a similar context, Barker et al. (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 was derived from a set of retrieved documents by a process known as subsumption. 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.

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.

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.

## 6.3 Material

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 a version of WordNet (Miller, 1990) were used for our experiment. Document frequency was determined using Google (at the time of experiment 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 6.1). Document frequency for a synset was estimated by averaging the document frequency of each member term.

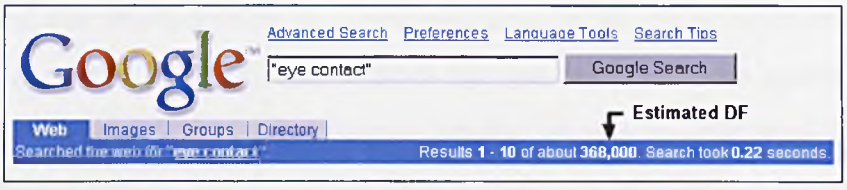



**Figure 6.1:** *Sample query in Google (Submitting a term eye contact with quotations. The estimated number of URLs containing the term was 368,000 and this would be the document frequency.)*

### 6.3.1 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.

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 prevailing 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 creators of WordNet used two ways to determine the frequency of occurrence of a word sense: first if the word occurred in the Sense Eval corpus, the commonest sense was measured from there; if the word did not occur in that corpus, then the commonest sense was determined by the WordNet creators based on their lexical/world knowledge (Miller, 1995).

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 occur-

rences, and that the definition of commonest sense in WordNet was correct in our corpus. The latter assumption is perhaps the one that should be questioned the most, after all, what if the corpus being examined 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 assumption is unlikely, experimental results presented in Section 6.4 appear to support it and therefore support the experimental design.

### 6.3.2 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 shown below.

```
act, human action, human activity (Top of hierarchy)
    action
        interaction
            contact
                eye contact (Bottom of hierarchy)
```

A total of 59,920 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. In order to address the disambiguation issue discussed in the previous section, 25,424 hypernym chains that consist of only the commonest sense synsets were used. Table 6.1 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

**Table 6.1:** *Length and distribution of hypernym chains*

| Chain length | Original | Disambiguated |
|---|---|---|
| 2 | 38 | 0 |
| 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 |

were used as the basic data in our experiment. Although the use of the disambiguated chains can eliminate the diversity of concepts provided by WordNet in the evaluation, the accurate estimation and distribution of the document frequency across all senses is not a trivial task, and it is beyond the scope of this thesis.

Finally, note that the level of specificity does not necessarily correlate with the length of chains in our data set. In other words, a term at the bottom level of a length 6 chain does not mean to be more general than the bottom level of a length 9 chain. The length of chains is likely to be determined by our knowledge about the diversity of categories to which a term belongs. Similarly, the difference between the level 2 and 3 in a length 6 chain is not always comparable with the difference between the same pair of levels in a length 9 chain.

## 6.4 Experiments

A three part experiment was undertaken to investigate the relationship between document frequency and term specificity, which are discussed below.

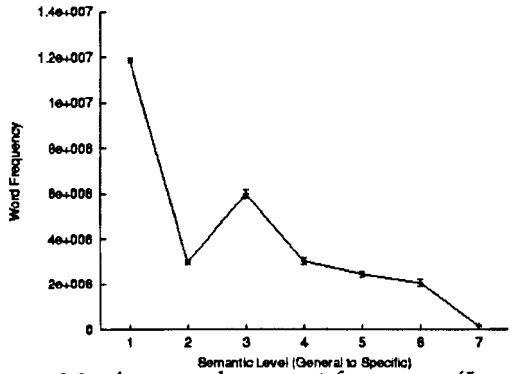



**Figure 6.2:** *Average document frequency (Length 7)*

### 6.4.1 Average document frequency

In 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 6.2.

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 of chains except length 3 and 4 (See Figure 6.3).

As can be seen, the shape 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, which were length 5, 6, 7, and 8, and consisted 73% of the hypernym chains in the disambiguated set. We will discuss the implication of the peaks in Section 6.5.

### 6.4.2 Parent-child pairs

The first 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

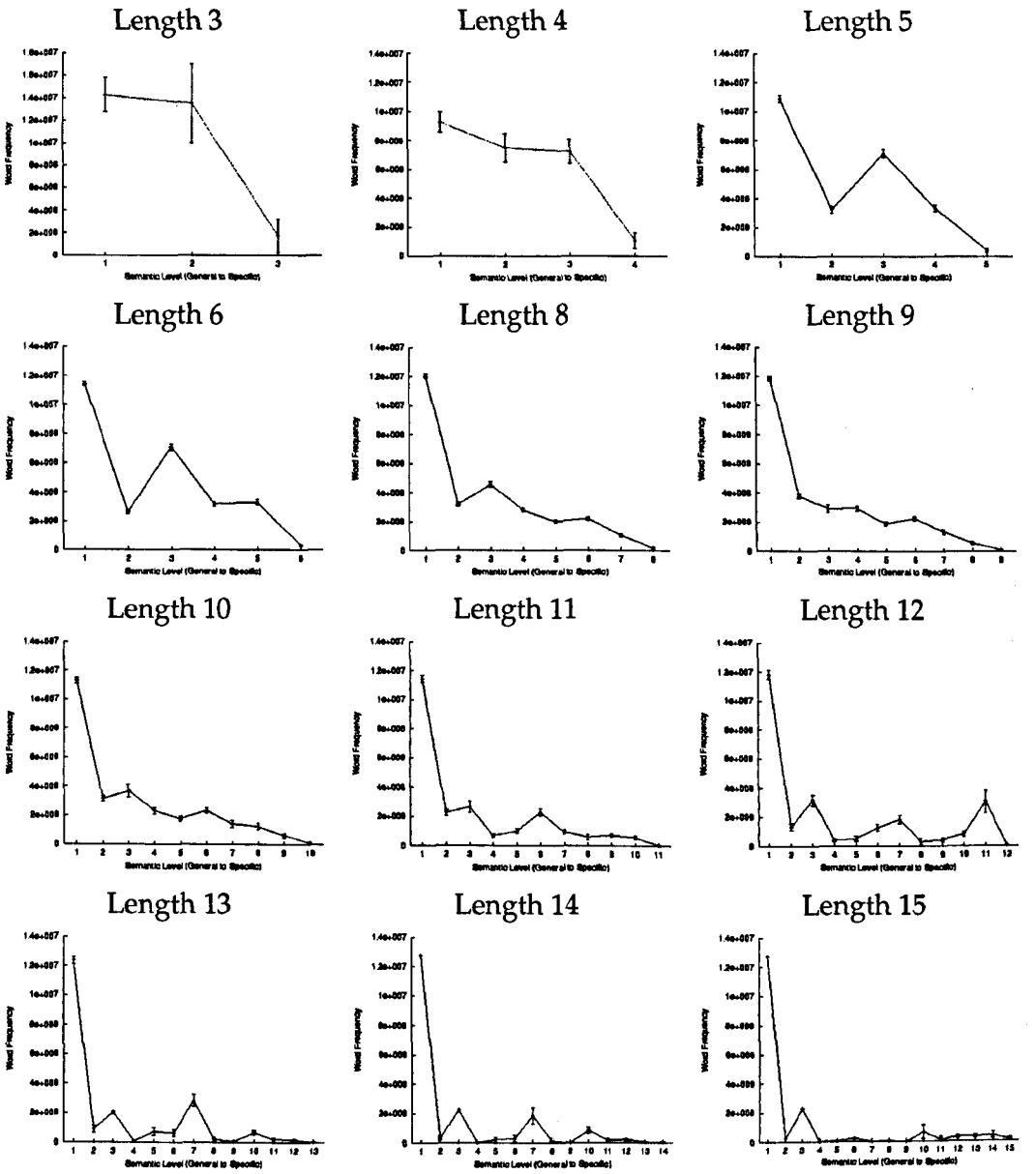



**Figure 6.3:** *Average document frequency: Length 3 to 15*

than their children do. It was also anticipated that this experiment should reveal the ability of document frequency to identify a parent term for a given pair. The result is shown in Table 6.2.

**Table 6.2:** *The rates for parents to have a higher DF (Length five to nine). The figures in brackets are the number of unique pairs between a parent and child.*

| Chain length | No of chain | Parent 1 Child 2 | 2 3 | 3 4 | 4 5 | 5 6 | 6 7 | 7 8 | 8 9 | Avg. (%) |
|---|---|---|---|---|---|---|---|---|---|---|
| 5 | 2696 | 56.5 | 61.0 | 65.0 | **82.4** | | | | | 66.2 |
| | | (23) | (123) | (486) | (2170) | | | | | |
| 6 | 5297 | 61.1 | 57.0 | 66.8 | 65.1 | **78.5** | | | | 65.7 |
| | | (18) | (79) | (298) | (910) | (4142) | | | | |
| 7 | 5681 | 60.0 | 58.0 | 67.5 | 60.7 | 69.0 | **81.8** | | | 66.1 |
| | | (15) | (50) | (163) | (417) | (1269) | (4532) | | | |
| 8 | 4793 | 61.5 | 60.0 | 67.0 | 57.5 | 62.4 | 68.4 | **78.3** | | 65.0 |
| | | (13) | (30) | (97) | (201) | (471) | (1185) | (4033) | | |
| 9 | 2773 | 60.0 | 72.7 | 62.3 | 59.3 | 55.3 | 65.4 | 60.4 | **79.8** | 64.4 |
| | | (10) | (22) | (53) | (91) | (190) | (353) | (808) | (2454) | |

There are two points to be emphasised in the table. Firstly, the number of unique 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 usually found in the lowest two levels. This indicated that document frequency was most reliable to identify a more general (or specific) term from a given pair if both terms were very specific. This echoed the result of Caraballo and Charniak's work.

An overall performance of this task for all 26,678 unique pairs found in our data set was 75.8%. This accuracy is respectable given that we only used document frequency data.

### 6.4.3 Effect of co-occurrence information

Up to this point, term specificity was determined through frequency of occurrence alone. The last part of our experiment examines the effect of co-occurrence information on the accuracy of term specificity estimation. Co-occurrence of terms has been exploited in the context of selecting query expansion terms (Peat and Willet, 1991;

Xu and Croft, 1996), and 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, 1997) and TREC-7 (Voorhees and Harman, 1998), namely, Topics 301-400. For each query, a set of the top ranked 500 documents was retrieved by a probabilistic IR system from the Financial Times (1991-94), LA Times (1989-89) and Wall Street Journal (1987-92) collections provided by TREC. As a result, 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.

We split the TREC document collection because we were interested in investigating the effect of co-occurrence information in multiple corpora. For example, the Financial Times and Wall Street Journal were likely to cover the financial issues more frequently than the LA Times. The period of covered articles was also different across the newspapers. These factors might have an effect on the range and frequency of vocabulary used in the corpora. As shown later in this section, the portion of the WordNet vocabulary found in the three corpora was different.

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 were recorded for 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. The result is shown in Table 6.3.

There are several points to be made.

- The document frequency obtained from a larger size of corpus was found

**Table 6.3:** *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 | 72.0% | 72.5% | 72.5% | 83.4% (+15.9) | 84.3% (+16.2) | 83.4% (+14.9) |
| TREC DF | 70.6% | 71.3% | 70.8% | 83.2% (+17.8) | 84.0% (+17.9) | 83.4% (+18.7) |
| Top500 DF | 65.1% | 67.1% | 66.1% | 82.0% (+25.8) | 83.2% (+23.8) | 81.4% (+23.1) |

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 more accurate 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 collection. 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 6.3.1, we were aware of the potential problem of WordNet's definition of commonest 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 6.4. As can be seen, the coverage of Google covers almost all WordNet vocabulary while the TREC collections covers just over 50%. 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. This result also highlights the usefulness of a large corpus such as the Web to determine term specificity.

**Table 6.4:** *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%) |

## 6.5 Discussion

Our findings have several implications to the relationship between the document frequency and term specificity. The first aspect is the shape of chains shown in Section 6.4.1. While there was a difference in the number of peaks across the chains, they all looked somehow similar. This partially echoes our intuition, that is, terms that are more specific generally have a lower document frequency (DF). Since the document frequency used in our experiment was obtained from the index of a heterogeneous web collection, it was unlikely that the shape of chains was significantly influenced by a particular domain or style of writing found in the collection. A stronger factor to form the observed shape appeared to be the fact that the chains had the highest frequency at the most general level and lowest frequency at the most specific level. However, the shape did not form a monotonically decreasing line, either. We found one or two peaks in the shapes. A longer chain tended to have more peaks. In the 1972 paper, Sparck Jones (1972) commented that the type of specificity defined by IDF was not necessarily the same as a semantic perspective. The shape of chains found in our results appears to support her speculation.

A prominent question to address is perhaps the cause of the peaks. The hypernym chains used in our experiment represented a series of associated concepts in

order of their specificity. The order of specificity was determined by the creator of WordNet based on (semi) formal classifications. Therefore, a shape of hypernym chains can be seen as a measure of correlation between the semantic specificity and frequency of occurrence in texts. A monotonically decreasing line can be found if they are always correlated. Therefore, the peaks observed in the middle level(s) suggest that there is a mismatch between the order of specificity and order of frequency. More specifically, people appear to use the concepts at a particular level more frequently than the other levels. Such a frequent level is known as *basic level* (Rosch, 1978).

There is a large body of research on human's cognitive categorisation process and basic level in Psychology and Cognitive Science[1]. For example, Rosch (1978) and Lakoff (1987), who were influenced by Wittgenstein's work on categorisation (Wittgenstein, 1953), investigated the effects produced by basic level categories. They found that people tended to acquire a particular level of abstraction (i.e., basic level) faster than the other levels. Also, they discuss that children's categorisation appears to be more accurate at the basic level compared to other levels. Examples of the terms at basic level are a *chair* or *dog* (Lakoff, 1987). The basic level effects indicate that the terms at the basic level are likely to be used more frequently than its parent (*furniture* or *animal*) or its child (*rocker* or *retriever*) (Rosch, 1978). Therefore, we can speculate that some of the peaks found in the hypernym chains are of the basic level.

**Table 6.5:** *Basic level of categories and document frequency*

| | Term | DF | Term | DF |
|---|---|---|---|---|
| Parent level | furniture | 3710000 | animal | 6570000 |
| Basic level | chair | 5280000 | dog | 7140000 |
| Child level | rocker | 436000 | retriever | 275000 |

Table 6.5 shows the examples of the basic level terms along with the DF data used in our experiment. As can be seen, the terms at the basic level had the highest DF among the three levels; the parent level has a higher DF than the child level; and the DF between the parent level and basic level varied across the domains.

[1]See Chapter 2 of Lakoff (1987) for an overview of the studies related to human's cognitive categorisation process in these fields.

It would thus appear that the notion of the basic level is helpful in providing an interpretation for the shape of the hypernym chains observed in our experiment.

The relationship between the basic level terms and their high DF values also seems to help us gain further insight into query formulation and expansion. For example, Jansen et al. (1998) reported that the searchers on the web tended to submit a short query with broad sense. This indicates that the terms at the basic level can be used more frequently in queries. Therefore, the query expansion based on a thesaurus (see Section 3.2.4) might benefit from the analysis of their queries from the basic level point of view. For example, if an initial query term belongs to the basic level, both its broader term (BT) and narrower term (NT) might have a lower DF value, thus, higher IDF weight. On the other hand, when the initial term is a NT of the basic level, query expansion based on the BT might have a higher DF value (lower IDF weight). This difference might affect the improvement of recall and precision in their application.

The second aspect is the accuracy of specificity estimation across the semantic levels. In particular, the result presented in Section 6.4.2 appears to somewhat contradicts the results presented by Caraballo and Charniak (1999) where they stated that frequency could determine specificity 86.3% of the time for parent-child pairs below the basic level and 45.5% above the level. In our result with the substantially larger multiple data sets than theirs, it would appear that the high levels of accuracy are really only apparent for the parent-child pairs found at the very bottom of a hypernym chain. For all other pairs regardless of the basic level, the accuracy is lower but relatively constant. Also, the accuracy rarely goes below 50% in our result. This indicates that while an overall accuracy of DF might be lower than term frequency used in their experiment, the performance of the former is likely to be more robust than the latter across the range of semantic levels.

Finally, our result in Section 6.4.3 suggests that the effect of co-occurrence information in the estimation of term specificity should not be underestimated. While the accuracy of estimating term specificity was found to correlate with the size of the corpus used to obtain the DF, the co-occurrence information can improve the estimation accuracy even with a small corpus. Therefore, our results provide an empirical evidence to support the use of co-occurrence information in the auto-

matic generation of concept hierarchies based on a set of retrieved documents (e.g., Sanderson and Croft, 1999).

It should be noted that other methods can also be used to improve the accuracy of term specificity estimation. Researchers in Computational Linguistics and Semantic Web are investigating the techniques based on the natural language processing and other knowledge-based resources to determine term specificity (e.g., Caraballo and Charniak, 1999; Cederberg and Widdows, 2003; Ryu and Choi, 2006). For example, Caraballo and Charniak (1999) considered the part-of-speech (POS) of words in the estimation of specificity. Their result suggests that the accuracy of estimation can be better at above the basic level when the POS tags were available in the collection, compared to a purely frequency-based method. Cederberg and Widdows (2003) applied a latent-semantic analysis to the extraction of hypernym concepts. Ryu and Choi (2006) tested several combinations of term frequency, document frequency, and POS information using the hypernym chains of a medical domain thesaurus. Their result suggests that the best combination can achieve over 80% of accuracy in the estimation. It would be interesting to investigate the performance of DF obtained from the web to a domain-specific collection, thus, this forms one of our future work.

## 6.6 Summary

This chapter 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 investigations 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.

From the first part of the 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 formed. More often, one or more middle levels in the chains held the second (or third) highest document frequency in the chains. We speculated

that this was due to the basic level effects.

The second part of the experiment focused on parent-child pairs in the chains to reveal the 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 75.8%, over three quarter of the cases.

The last part of the 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.

When the document frequency is used as a means of generating a concept hierarchy, therefore, the frequency taken from a larger collection is likely more accurate than the one from a smaller set. A wider range of vocabulary will also be found in the larger collection. However, there will be the case where it is infeasible to access to the data set as large as the index of a major search engine's collection. In such a case, the co-occurrence information was shown to be useful for improving the accuracy of ordering concepts based on specificity.

A limitation of the work presented in this chapter is due to the assumption applied in the creation of our data set, that is, the default sense provided by WordNet will be used in the majority of occurrence of concepts in the document collection. Therefore, our findings might not be valid when a domain-specific document collection is studied. Future work in this area would be to examine to what extent the statistics obtained from a large collection can determine the most popular sense of concepts in a domain-specific application.

The next chapter will explore a lexical aspect of concepts as an alternative approach to generating a concept hierarchy.

# Chapter 7

# Experiment II: Extracting parent-child descriptions

## 7.1 Introduction

In Chapter 5, we discussed two major aspects of words which can be explored as a means of determining the specificity of concepts. The main motivation was to derive a hierarchical structure of concepts from a set of retrieved documents to support a user's query reformulation and searching in an interactive fashion. The previous chapter presented the evaluation of the statistical aspect of concepts. This chapter presents a study which explores a lexical aspect of concepts.

The study in this chapter is based on a system, the Descriptive Phrase Finder (DPF), that retrieves descriptions of a query term from free-text, and reports a large scale experiment 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 the query *Microsoft*. The system only uses simple pattern matching to detect a description, and ranks the sentences that hold the descriptive phrases based on a mixture of evidences such as the frequency of occurrence, the accuracy of patterns, and document location. 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.

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 effect was demonstrated in the results of the Very Large Collection (VLC) track of TREC-6 (Hawking and Thistlewaite, 1997). In the track, the same topics as those used in the Ad Hoc task were applied to a 20Gb collection, which was 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 there were simply more relevant documents that held a high percentage of the specified query terms in the VLC collection. The design of the DPF system was motivated by the result from the VLC track. In other words, we focused on a relatively simple approach to extracting descriptive phrases from a large amount of texts. The simplicity of our approach would also enable the DPF system to be used in a domain-independent application.

The rest of this chapter is structured as follows. The next section presents the details of the DPF system. The types of patterns used, other components for ranking, and relevance assessments for the evaluation are discussed. Then the result and analysis of the experiment are shown. The experiment looks at the effect of collection size used to extract descriptive phrases, as well as a form of precision and recall to evaluate the effectiveness of the DPF system. The last sections discuss the findings of the experiment and relation to other studies, followed by a summary of this chapter.

## 7.2 Materials

The details of the system and evaluation methodology used in exploring the lexical aspects of concepts are as follows.

**Table 7.1:** *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* | ... broken bones **and other** injuries ... |
| especially | *qn* (**especially** \| **including** ) *dp* | ... injuries **especially** broken 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.

### 7.2.1 Descriptive Phrase Finder

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 7.1.

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 (Radev, 1998), 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 before the

experiment and is described in Section 7.2.2. The key phrase patterns shown in Table 7.1 were based on the literature (e.g., Hearst, 1998; Radev, 1998). Some of the variant patterns were populated during a preliminary testing of the patterns using our training document set (see Section 7.2.2) as well as a testing with a search engine.

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. Aspects of this approach are to be found in QA systems utilising redundancy (Clarke et al., 2001).

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.

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 $a$, $b$, $c$, $d$ are the tuning constants which are discussed below.

### 7.2.2 System tuning and relevance assessment

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.

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 training set was assessed to measure the accuracy of the patterns.

The relevance of sentences was assessed by the author of this thesis. In order to reduce subjective bias in the relevance judgements, a judgement scheme similar to TREC Ad-Hoc track[1] was used during the relevance assessment. More specifically, a sentence was judged as relevant when a descriptive phrase contained some information to help a user understand more about the query s/he submitted. Therefore, the assessment was a binary judgement. While this might not entirely eliminate subjective bias in the judgements, it was at least complemented by the consistency of criteria applied to the relevance assessment.

As with all relevance judgements, however, there were some sentences that were hard to decide on. For example, in one sentence containing the query "Adolf Hitler", he was described as a person, "... not only for Kraft but for people such as Adolf Hitler and Adolf Eichmann ...". Although this was 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.

It should also be noted that judging relevance was on sentences and not on extracted descriptions or other units 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. The same assessment approach was taken in the

[1]See `http://trec.nist.gov/data/reljudge_eng.html` [Accessed: 6/1/07].

**Table 7.2:** *Accuracy of key phrase patterns*

| | Code | Not Relevant | Relevant | Total | Accuracy (%) |
|---|---|---|---|---|---|
| No pattern | na | 6,424 | 872 | 7,296 | 12 |
| especially | es | 0 | 0 | 0 | 0 |
| qn, dp, | ap | 89 | 63 | 152 | 41.4 |
| is a | ia | 23 | 18 | 41 | 43.9 |
| including | in | 20 | 17 | 37 | 45.9 |
| or other | oa | 1 | 1 | 2 | 50 |
| such as | sa | 59 | 59 | 118 | 50 |
| Acronym | ac | 14 | 23 | 37 | 62.2 |
| and other | ao | 9 | 23 | 32 | 71.9 |

experiments presented in the following section.

The result of the accuracy and coverage based on the training set is shown in Table 7.2. 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. This result and other informal tests were carried out to determine the value of the parameters $a$, $b$, $c$, $d$ (as defined in Section 7.2.1), and they were set to 2000, 1, 75, and 500, respectively. 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.

A sample output of the DPF system used in the third experiment is presented in Table 7.3. From the left side, each column shows a rank, final score, matched pattern code (c.f. the 2nd column of Table 7.2), 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.

## 7.3 Results and analysis

Our evaluation of the DPF system is based on a set of proper nouns (or named entities). The motivation for using proper nouns in the evaluation is as follows. Firstly, the proper nouns are often not available from the conventional resources such a dic-

**Table 7.3:** *A sample result of the DPF system (Query: tony blair)*

| Rank | Score | Code | KP | WC | SL |
|---|---|---|---|---|---|
| 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 | 8.860 | 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 | 8.750 | 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 | 8.750 | 0 | 1 |
| | **Tony Blair** is an Alien! | | | | |

tionary, thus, their information is worth extracting from texts. Secondly, people are more likely to find proper nouns unfamiliar or too specific in query expansion, than the common words found in a dictionary. As we discussed in Section 3.4, the lack of knowledge or information on suggested terms appears to cause a difficulty in selecting expansion terms (Belkin et al., 2003). Therefore, it seems worth investigating the effectiveness of the DPF system using proper nouns.

The document collection was taken from the TREC LA Times test collection. While there were several proper nouns found in the TREC topics which used LA Times collection, it was decided to create a new set of queries. One reason for this was that it was not always clear if the creators of the TREC topics intended to learn more about the proper nouns used in the topic descriptions. Another motivation was that we were interested in including the information needs expressed by real users in the experiment. For this, we asked the member of the Department of Information Studies about the words which they were (or used to be) interested in learning more about them.

Consequently, 76 queries were suggested by the author and colleagues, of which 10 were not present in the collection and a further 16 that only occurred a small number of times ($N < 20$). These 16 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 a small number of sentences retrieved by those queries. The remaining 50 queries were used in the experiment. A total of 16,111 candidate sentences were found in the test set of collection for the 50 queries, and they were evaluated for relevance in the same manner described in Section 7.2.2.

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.

### 7.3.1 Collection size

In this part of the experiment, random samples of the test set were taken and the effectiveness of the key phrase pattern matcher was measured for each of these

**Table 7.4:** *Effectiveness of the phrases across the collection size*

| | 100% | 75% | 50% | 25% | 10% |
|---|---|---|---|---|---|
| Precision at 1 | 0.75 | 0.78 | 0.69 | 0.63 | 0.62 |
| Precision at 5 | 0.51 | 0.51 | 0.46 | 0.38 | 0.32 |
| Precision at 10 | 0.42 | 0.40 | 0.35 | 0.28 | 0.24 |

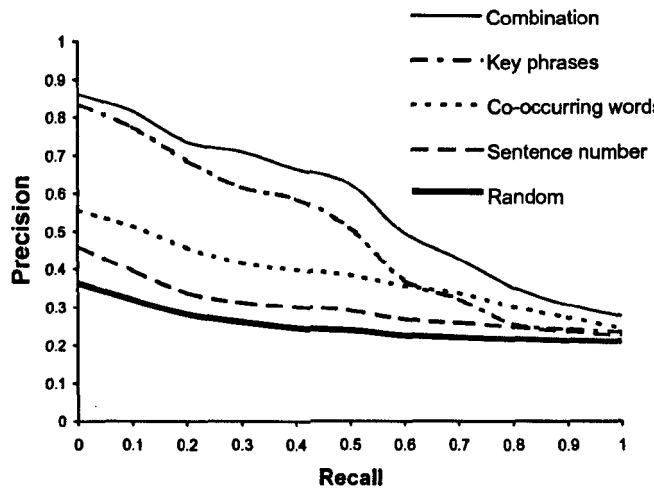



**Figure 7.1:** *Recall and precision of the DPF system*

samples. Results from this experiment are shown in Table 7.4. 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 due to the decreasing likelihood of the system finding a sentence holding one of the key phrases. This result echoes the results reported by the VLC track discussed in Section 7.1.

### 7.3.2 Recall and precision

The effectiveness of each of the individual sentence ranking criteria was tested along with combination formula derived above. We also devised a random baseline for a comparison. 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 recall-precision graph was plotted showing the effectiveness of the four strategies plus random retrieval (see Figure 7.1). As can be seen, the three criteria and their combination do better than random retrieval. A sentence ranking based purely

**Table 7.5:** *Performance of ranking factors by precision*

| 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 |

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 for recall levels 0.1 and 0.2 and significant at for all higher values of recall. When evaluated with precision oriented measures a similar picture emerged.

Table 7.5 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 ($p \leq .01$). 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 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%).

### 7.3.3 Sample descriptive phrases

Table 7.6 shows some of the descriptive phrases found in the top ranked sentences. As can be seen, while the length of descriptive phrases can vary, in many cases, the phrases contain the words that can be used to form a parent-child relationship with

**Table 7.6:** *Top ranked descriptive phrases*

| Query | Descriptive phrase(s) |
|---|---|
| AIDS | a life-threatening disease; the worst thing to happen in the 20th Century; a human disaster; a virus infection; acquired immune deficiency syndrome |
| Bob Dylan | giants; artists |
| Diane Sawyer | reporters; media and fashion stars |
| Hitachi | Japanese semiconductor giants; large Japanese electronics companies; industrial powerhouses |
| Microsoft | The world's largest PC software publishing house; software companies |
| NATO | the National Association of Theater Owners; North Atlantic Treaty Organization; international organization |
| Nike | companies; the nation's top sneaker firm |
| Star Wars | strategic programs in fiscal 1991; high-technology weaponry |
| UNIX | developed operating systems |

the queries.

To illustrate how these descriptive phrases can be used to form a hierarchy, we take *Hitachi* and *Microsoft* for instance, and assume that the descriptive phrases of the two names are found in the retrieved documents. Both names have the word "companies", thus, this can be used as their parent concept. However, since Hitachi has a phrase "electronics" companies and Microsoft has a phrase "software" companies, the parent concept, companies, can further be divided into two children as shown in Figure 7.2. In this way, the hierarchy still has a possibility of adding other types of companies found in texts. Similarly, other (electronics or software) companies can also populate the hierarchy based on the common phrase (such as electronics or software). More details of the hierarchy creation based on the key phrase will be discussed in the next chapter.

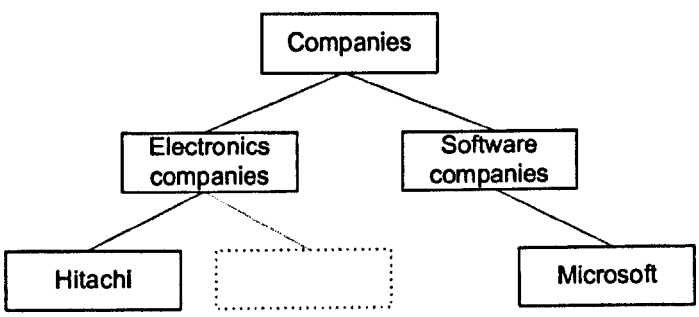



**Figure 7.2:** *Sample hierarchy based on descriptive phrases.*

It is also interesting to point out that more than one sense can also be detected for an ambiguous concept by our system (e.g., NATO).

## 7.4 Related Work

Since the completion of the work presented in this chapter, there have been some relevant studies carried out by researchers. For example, Fujii and Ishikawa (2000, 2001) have developed a system that had a similar aim to the DPF system. Their system aimed to extract encyclopaedic knowledge of technical terms from the Web. It was 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. One of their features which were not exploited in our system was the structure of documents defined by the HTML tags. Their system was designed to analyse a certain set of HTML tags (e.g., <DD>, <DT>, <OL>, and <UL>) to extract the encyclopaedic knowledge. The candidate descriptions were then classified 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 domain, descriptions were ranked based on an N-gram probability. Three top ranked descriptions from each domain were the output of the system.

They reported that their system retrieved at least one relevant description in the top 3 results for 90% of 85 queries. Our system retrieved at least one relevant description in the top 5 sentences for 90% of 50 queries. One of the factors affecting the difference of performance appears to be the size of collection used to extract the description. In Fujii and Ishikawa's work, the web document collection, which was larger than our LA Times collection, was used via a search engine. As we discuss below, our follow-up experiment suggests that the performance of the DPF system can be comparable to their system when a larger document collection was available.

An advantage of their system appears to be a disambiguation process based on a classification technique. When a query has several senses or the system retrieves similar descriptions, their system appears to be better at grouping the descriptions than our approach. While our results also suggest that the descriptions for multiple

senses of a query can be extracted by the DPF system, we did not attempt to group them into some structure. However, the success of their grouping seems to depend on the pre-defined 19 domains. Our system, on the other hand, was designed to work without the restriction of domains.

A follow-up experiment of the DPF system was carried out by an MSc student who was jointly supervised by Dr. Sanderson and the author of this thesis. Given that the results of the experiment presented in this chapter 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 follow-up experiment was designed to evaluate the effectiveness of the DPF system with an larger collection, the web, via Google (Joho et al., 2001). 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 follow-up experiment, and a more stringent form of relevance was devised. Like the previous experiment, we asked the colleagues to suggest potential queries. However, this time we also asked sample answers of each query: for example, "the Prime Minister of Great Britain" for the query *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 descriptive phrases was not enough for this aim. Only a descriptive phrase 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. This relevance judgement scheme was referred as *stringent relevance* while the one used in this chapter was referred as *binary relevance.*

A total of 50 queries were tested with the stringent relevance 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. 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

**Table 7.7:** *Percentages of successfully answered queries: Comparison with the previous experiment (LA Times, N=50)*

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

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.

As a means of comparison, a repeat of the previous experiment was also conducted using the binary relevance scheme 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 previous experiment is illustrated in Table 7.7. This shows that an improvement over the previous experiment was achieved by moving to a larger collection.

These results show that the DPF searching the Web works better than the previous experiment using LA Times, thus, 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 experiment, 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 contribution of the statistics increases, thus, the performance of the system was improved.

## 7.5 Summary

In this chapter, we presented a system that retrieved descriptive information about a query term, and evaluated the effectiveness using proper nouns. The design of the system was partly motivated by findings from the VLC track of TREC (Hawking and Thistlewaite, 1997), 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 chapter also described two criteria for relevant judgement of retrieved descriptive phrases: binary and stringent relevance. The latter relevance criterion using the key answers was used to evaluate the effectiveness of the system to retrieve highly relevant descriptive phrases. In the binary relevance, the system retrieved at least one relevant description in the top 5 sentences for 90% of the tested queries, while with the stringent relevance, the system retrieved at least one highly relevant description in the top 20 sentences for 88% of the queries. When the size of collection was increased to the Web, the performance was further improved. Therefore, this seems to justify our approach of using a relatively simple technique with a large scale collection.

In Section 7.3.3, we illustrated how a hierarchy can be formed using the descriptive phrases. We speculate that the relationship found by the lexically-oriented technique can be an interesting alternative to form a hierarchical structure of concepts. We also speculate that the association between concepts might be more intuitive in the lexical technique than the statistical approach due to the linguistic contexts captured in the former technique.

These two different approaches to derive a hierarchical structure of concepts from texts will be implemented, visualised, and finally integrated into an interface of an IR system in the next chapter. We will study the interaction between the users and our concept hierarchies which will be designed to assist their query reformulation and searching.

One of the limitations of the study presented in this chapter would be the range of query tested in the experiment. We mainly focused on proper nouns in the evaluation because we speculated that they were less likely to be found in a dictionary, thus, the searchers appeared to benefit from additional information on proper nouns compared to the common words. However, the presence of descriptive phrases might depend on the intended readers of texts. For example, the word *UNIX* might not have much description in a programming tutorial text. Therefore, while our approach to extracting parent-child descriptions was intended to be domain-independent, the range of descriptions which can be extracted from texts might vary when the system was applied to a domain-specific document collection.

# Chapter 8

# Experiment III: User interaction with a concept-based query reformulation supporting system

## 8.1 Introduction

The previous chapters investigated the different approaches to automatically generating a concept hierarchy. The main objective of these hierarchies was to generate a summary of retrieved documents based on the hierarchy, and to support a searcher's query reformulation and searching. This chapter presents the user study of a search interface which incorporates the concept hierarchies into the interface of an IR system.

This chapter is structured as follows. Firstly, our specific research hypotheses investigated in this study are stated. Secondly, the experimental design adapted for our user study is discussed. The details of the experimental systems are then described. The analysis of the experimental results is presented, followed by the discussion on the implications of the results.

## 8.2 Hypothesis

Our overall research hypothesis investigated in this chapter was that the concept-based interactive query expansion tool could facilitate a user's interactions with an IR system. To investigate this hypothesis, specific sub-hypotheses were devised and grouped into the following aspects of interactive search. The hypotheses were stated in contrast to an existing search interface which would offer the basic search functions such as submitting a query, presenting the search result, and accessing to the full-text of retrieved documents.

**Query reformulation:** The concept hierarchy was designed to present a set of the candidate expansion terms extracted from a set of retrieved documents. Therefore, we assumed that a user's query reformulation could be assisted. To investigate this aspect, the following hypotheses were considered.

- [*H*.1] Concept hierarchies reduce the amount of effort required for manual query reformulation to complete a task.
- [*H*.2] Concept hierarchies diversify the range of search vocabulary employed to complete a task.

**Browsing of retrieved documents:** Since the search interface allowed a user to browse a subset of retrieved documents by accessing the terms in the hierarchy, we assumed that a user's browsing of retrieved documents could be facilitated. To investigate this aspect, the following hypotheses were considered.

- [*H*.3] Concept hierarchies encourage participants to view the documents that are retrieved in a low rank.
- [*H*.4] Concept hierarchies improve the successful click-through rate (a portion of click-through documents in which perceived relevant information was found).

**Task perception and performance:** If the query reformulation and search result browsing were successfully facilitated, a user's task performance could be improved. In particular, since the hierarchy was designed to offer a more control on the browsing of retrieved documents, the precision was expected to improve. Also, a user's perception on the task should be improved. To investigate this aspect, the following hypotheses were considered.

- [*H*.5] Concept hierarchies improve the task performance, especially in precision.
- [*H*.6] Concept hierarchies enable participants to perceive that the searches are easy and successful.

**Information seeking behaviour:** Access to the concept hierarchy could be seen as an indication of the need of support during a search task. Since a user might require different support at a different stage of a search session, the following hypotheses were considered.

- [*H*.7] The frequency of the access to the hierarchies changes over the stage of a search session.
- [*H*.8] The depth of selected concepts in the hierarchies changes over the stage of a search session.
- [*H*.9] The frequency of the access to the hierarchies correlates with users' perceptions of search topics.

**Post-search learning effect:** Finally, since the concept hierarchies were designed to offer an overview of retrieved documents during a task, we assumed that a user would increase the knowledge of a search topic after a search session. To investigate this aspect, the following hypotheses were considered.

- [*H*.10] Participants can generate an improved query for a topic after a search session.

- [$H$.11] Concept hierarchies can improve participants' perception on the learning of a search topic after a search session.

The next section presents the design of the experiment which examines these hypotheses.

## 8.3 Experimental design

As can be seen from the range of hypotheses proposed in the previous section, this study was interested in a holistic approach to evaluate the effectiveness of the automatically generated concept hierarchies in interactive IR. In other words, user interaction with a search interface, subjective perceptions on the system and task, and information seeking behaviour were considered as important to investigate as classic measures based on precision and recall. One of the IR test collections which were designed to facilitate such a study was the Interactive Track in TREC (Over, 1997, 1998; Hersh and Over, 1999). This section first describes the Interactive Track approach, followed by our adaptation of the approach in the current study. The characteristics of participants and experimental procedure taken are then discussed.

### 8.3.1 TREC Interactive Track approach

The traditional evaluation of IR systems, such as TREC AdHoc Task, has been carried out in the absence of real searchers. The traditional approach is effective when researchers are interested in benchmarking the retrieval effectiveness of different ranking algorithms, in estimating an underlying model's parameters, and other system-centric aspects of IR systems. However, the traditional evaluation based on the precision/recall measure is only one way to improve the performance of IR systems. Another approach is to study real users' interaction with a search interface from the aspects such as query reformulation, browsing of retrieved results, and/or behavioural changes in the search session, topic familiarity, or task complexity (Borlund, 2000; Beaulieu, 2000; Efthimiadis, 2000; Hersh et al., 2001; Shiri and Revie, 2003; White et al., 2005).

The Interactive Track of TREC was developed to support user-oriented studies with an interactive IR system. The Track was designed to focus on the searchers rather than the systems, and on behavioural details rather than summary measures by precision and recall. The Track shared the advantages of traditional test collections such as the common set of documents, topics, and relevance judgements used by the Track participants. However, the Track offered a different search task to facilitate a searcher's interaction with the search interface. The task adapted in our study was called the *instance finding task* which was developed in the Track over the years (Over, 1997, 1998; Hersh and Over, 1999).

While the Ad Hoc task assumed that a searcher was looking for as many documents relevant to the topic as possible, the instance finding task assumed that the searcher was looking for as many *different* instances relevant to the topic as possible. Therefore, finding duplicate relevant information (instance) had no reward in the task. This difference has advantages. Firstly, the search situation assumed in the instance finding task was likely to increase the level of realism manipulated in the experiment. Secondly, since the searcher was asked to find the different aspects of a topic, it was likely to motivate the searcher to interact with the search interface more intensively than the Ad Hoc task. The example of the instance finding task is shown in Figure 8.1.

```
Title:
  tropical storms

Description:
  What tropical storms (hurricanes and typhoons) have
  caused property damage and\or loss of life?

Instances:
  In the time allotted, 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.
```

**Figure 8.1:** *Instance finding task description (Topic 408i).*

The task description was a modification of existing TREC topics. For example, Topic 408i shown above was based on Topic 408 in the TREC-8 Ad Hoc task. The six topics devised by the Track in TREC-8 were: tropical storms that caused property damage (408i), countries that imported Cuban sugar (414i), countries (except U.S. and China) that have had a declining birth rate (428i), developments and use of robotic technologies (431i), countries that have experienced an increase in tourism (438i), and finally, countries in which tourists have been subject to acts of violence (446i). The document collection to be searched was the Financial Times (1991-1994) which contained 210,158 documents (564MB).

In the Interactive Track, the precision/recall were measured in a different way from the Ad Hoc Track. The Interactive Track participants were asked to submit the documents in which subjects identified perceived relevant instances. A document pool was formulated from the submitted documents. The relevance assessments were carried out on the document pool and the NIST assessors identified the official relevant instances. Based on these data, the *instance recall* and *precision* were calculated for each topic as follows.

$$\text{Instance Recall} = \frac{\text{Number of relevant instances in submitted documents}}{\text{Total number of official relevant instances}} \tag{8.1}$$

$$\text{Instance Precision} = \frac{\text{Number of submitted documents with a relevant instance}}{\text{Total number of submitted documents}} \tag{8.2}$$

Note that the Track's definition of instance recall and precision was based on the *documents* submitted by participants and instances identified by the assessors. Therefore, the calculation of precision and recall was not based on the actual instances identified by the subjects of individual participated groups. Later, Wu et al. (2001) proposed the *true* instance recall and precision as follows.

$$\text{True Instance Recall} = \frac{\text{Number of relevant instances retrieved}}{\text{Total number of official relevant instances}} \tag{8.3}$$

$$\text{True Instance Precision} = \frac{\text{Number of relevant instances retrieved}}{\text{Total number of retrieved instances}} \tag{8.4}$$

As can be seen, Wu et al's definition was based on the actual instances identified by subjects of a user study. In this chapter, we use the true instance recall and precision since they are more intuitive to understand the task performance. Also, Wu et al. (2001) show that these two measures of recall and precision are significantly correlated in their study. As mentioned above, the true instance and recall were not determined in the Interactive Track because only documents were submitted by participants, and not specific instances. In our study, the retrieved instances were recorded during the experiment to measure the true instance recall and precision.

Apart from the basic components of a test collection, the Interactive Track also provided a core set of questionnaires. An entry questionnaire was used to establish participants' background information. A post-search questionnaire was used to capture participants' subjective assessments on the topics and searches. A post-test questionnaire was used to capture participants' overall feedback on the systems and other opinions.

### 8.3.2 Our design

As can be seen, the Interactive Track aimed to offer a framework for the holistic evaluation of an Interactive IR system, thus, we decided to use it as the basis of our evaluation. The instance finding task was also relevant to our overall research goal of investigating the effectiveness of the automatically generated concept hierarchies. On the one hand, the hierarchies were designed to support an effective browsing of retrieved documents by organising the extracted terms hierarchically. Finding different instances can benefit from focusing on multiple subsets of retrieved documents using the hierarchies. On the other hand, the hierarchies were designed to support query reformulation by increasing the awareness of the relationship among the query terms, expansion terms, and retrieved documents. Finding different instances might require an effective use and management of search vocabulary, thus, benefit from the range of related terms presented in the hierarchies. Finally, the Interactive Track approach allowed us to study the searcher's interaction with the hierarchies over a different stage of search.

In this study, we used the same set of documents, topics, and relevance judge-

**Table 8.1:** *System and topic rotation*

| Subject | System | Topic | System | Topic | System | Topic |
|---|---|---|---|---|---|---|
| 1 | 1 | 408i | 2 | 428i | 3 | 431i |
| 2 | 1 | 414i | 3 | 438i | 2 | 431i |
| 3 | 2 | 438i | 1 | 428i | 3 | 446i |
| 4 | 2 | 446i | 3 | 408i | 1 | 431i |
| 5 | 3 | 414i | 1 | 438i | 2 | 408i |
| 6 | 3 | 428i | 2 | 414i | 1 | 446i |
| 7 | 1 | 408i | 2 | 428i | 3 | 431i |
| 8 | 1 | 414i | 3 | 438i | 2 | 431i |
| 9 | 2 | 438i | 1 | 428i | 3 | 446i |
| 10 | 2 | 446i | 3 | 408i | 1 | 431i |
| 11 | 3 | 414i | 1 | 438i | 2 | 408i |
| 12 | 3 | 428i | 2 | 414i | 1 | 446i |

ments (i.e., official relevant instances) as the Interactive Track in TREC-8. However, some modification was made to the Track's design to investigate our research hypothesis. The Track guideline suggested a minimum of twelve subjects, six topics per subject, and two systems (one control and one experimental system) for the Track participants (Hersh and Over, 1999). For each subject, three topics were searched by one system and three topics were searched by another system. The order of topics and systems presented to subjects was rotated to reduce an interaction of subjects and topics in a performance measure of systems. However, the same system was used for the first three topics in a row followed by the other system for the rest of topics.

The changes made in our design were as follows. In our experiment, twelve subjects were recruited and each subject performed a task on three out of six topics using a different order of the three systems (one control and two experimental systems). The rotation of the systems and topics is shown in Table 8.1. System 1 was the control system and System 2 and 3 were the experimental systems. As can be seen, the order of the systems was rotated across participants so that each system was tested by the same number of participants and same frequency of order. The topics were allocated so that each system was tested by the same frequency of six topics at the end of experiment. However, all combinations of topic order were not tested in our study due to the large number of possibilities (i.e., there are 120

ways to select three from six topics in a different order). Therefore, we set a higher priority to protect an effect of system order than an effect of topic order, since the domain of topics appeared to be different. The exception might be Topic 438i and 446i where both were related to the tourism, but 438i focused on an increase of tourists while 446i focused on the violence to the tourists. In summary, the order of systems and frequency of topics tested by each system were balanced in our design, but the order of topics was not.

In the result section, we run the Analysis of Variance (ANOVA) test to see if there was an interaction effect between three independent variables: system, topic, and order. When a significant interaction effect was found between system and topic, it meant that a system's performance varied over topics, which was common in IR systems. When a significant interaction effect was found between system and order, a measure was influenced by the order of system presented to participants. When a significant interaction effect was found between topic and order, a measure varied over topics but it was influenced by the order of topics presented to participants.

Another difference from the original design was the amount of time given to a search. Our participants were given 10 minutes as opposed to 20 minutes in the Track. This decision was made based on the experimenter's general observation of a pilot study which was carried out to practice the experimental procedure and to find any system bugs. The pilot study was based on the Ad Hoc task and four representative test participants were given 20 minutes for each of three topics. We noticed that participants of the pilot study tended to become less active in finding relevant information towards the end of a whole session. One of participants also expressed that 20 minutes was too long to concentrate on search. Therefore, while the task was different, the time was shortened in the main experiment to ensure that participants can perform the searches with a similar level of engagement across the systems. After the end of the main experiment, we made an informal comparison with the Interactive Track runs. The average number of save documents in our study was 7.5 (SD: 3.8) compared to 8.5 (SD: 4.5) in the Rutgers University (Belkin et al., 1999), who had the largest number of subjects in the TREC-8 Interactive Track. The average number of retrieved instances in our study was 9.2 (SD: 5.1) compared to approximately 11.9 (SD not reported) in the Sheffield University (Beaulieu et al.,

1999), who used the same retrieval engine as used in our experiment. Therefore, there seemed to be some impact of the time difference in the task performance, but the difference was smaller than we had expected. However, it was possible that our participants were tested under a higher level of time pressure than a 20 minutes task.

Finally, we added a small task at the end of each search to investigate the effect of topic learning. We asked participants to generate an optimised query which would retrieve as many relevant instances based on their search experience. We called this small task the *query optimising task*. The objective of this task was to examine if participants' learning of topics can be detected from their optimised query. Section 8.5.6 presents the result of this task.

The next sections describe participants and procedure of the main experiment.

### 8.3.3 Participants

Twelve people were recruited for the user study. The recruitment was carried out by our call for participation distributed to the mailing lists of the Department of Information Studies and Department of Computer Science of the University of Sheffield. The entry questionnaire established that participants consisted of two females and ten males. Their occupation at the time of the study was Research Assistant (3), PhD students (8), and Assistant Professor (1) in either of the Departments. They had on average 6.7 years of online search experience (Min: 3, Max: 10, SD: 2.6). The distribution of age[1] was 18-27 (4), 28-37 (7), and 38-47 (1). The questionnaire also indicated that six used a search engine daily, four used at least once in a week, and two used at least once in two weeks. The frequent purposes for searching were related to work, entertainment, and shopping. Their favourite search engines (multiple answers) were Google (12), Yahoo (4), AltaVista (3), Ask Jeeves (1), and HotBot (1). Three of them had previously participated in a TREC experiment, but none had performed the instance finding task.

In summary, the subjects in this study did not necessarily represent the general population. Participants were likely to have advanced knowledge of search

[1]The entry questionnaire of the Interactive Track did not ask a specific age.

engines and other related information science and computer science technologies. We selected this population to get critical feedback on the systems since this was the first user study using our experimental systems. The further study on other populations is our future work.

### 8.3.4 Procedure

For each subject, the experiment was carried out in the following manner.

1. At arrival time, the motivation of the study was verbally informed. "We developed a tool to help users find documents and formulate a query, and we are interested in evaluating the tool using real users."

2. The subject was asked to complete an entry questionnaire. This contained the questions about participant such as the gender, age, and search experience.

3. The subject was explained the nature of instance finding task using an instruction (See Appendix A) and training task description. The instruction gave situational contexts of a task to participants using a simulated scenario (Borlund, 2000). The subject was reminded that the task aimed to find as many different instances as possible.

4. The subject was explained the nature of document collection used in the experiment. The subject was reminded that the database consisted of news articles in 1991-94.

5. The subject was given a training session with one of the experimental systems for 10 minutes based on the training topic. The subject was informed that the top 200 documents were available for each query, and the hierarchy was generated from the words and phrases found in those documents. The questions regarding the task and system were answered by the experimenter.

6. The subject was given the first topic and asked to start a search when s/he was ready. The subject was given 10 minutes to search. Before the start, the subject was reminded to write down the perceived relevant instances on a sheet of provided paper. The experimenter was monitoring the search to deal with a system problem.

7. At the end of the search, the subject was asked to formulate an optimised query for the topic.

8. The subject was then asked to complete a post-search questionnaire. This contained the questions about the subjective assessments on the search and

system. When one of the experimental systems was used, there were several additional questions to captures the assessments on the hierarchies.

9. Step 6 to 8 was repeated three times based on the rotation table shown in Table 8.1.

10. At the end of three topics, the subject was asked to complete a post-test questionnaire. This contained the questions about the system preference and other feedback on the systems.

The instruction and questionnaires were based on the ones provided by the TREC-8 Interactive Track. Questions were added to the experimental systems' post-search questionnaire to capture participants' assessments on the concept hierarchies. Changes were also made to the post-test questionnaire to list three systems as opposed to the original design of two systems. The instruction and questionnaires are found in Appendices.

The entire session lasted between 60 to 90 minutes. The participation was voluntary and no reward was given. The following section presents the detail of the systems evaluated in our study.

## 8.4 Systems

Three systems were created for the experiment. The first system (System 1) was a control system which had the basic search function such as submitting a query, viewing the full-text of retrieved documents, and moving to a next result page. The second and third systems (System 2 and System 3) were the experimental systems where the concept hierarchies were generated from a set of retrieved documents and incorporated into the interface of System 1. This section describes the details of the three systems.

### 8.4.1 Control system

The Okapi system (Robertson et al., 1997) was used as the back-end retrieval engine. The Okapi system indexed the document collection and performed retrieval in response to a query. The passage-based BM25 function was used to rank the


**Figure 8.2:** *A screenshot of Control system*

documents. The passage-based BM25 function is similar to the original BM25 function, but the ranking of documents was based on a passage-level as opposed to a document-level. Therefore, the document which had a higher scoring passage was ranked higher than the others were. This function allowed us to extract the best passage from the documents in query time, which was presented as a query-biased summary of documents in the search results. The length of passage was set to 50 words. The parameters of the function were taken from Robertson et al. (1998) where the optimised value was estimated for the TREC document collection.

The search interface (shown in Fig 8.2) was implemented by the Common Gateway Interface (CGI) and written in Perl and Javascript. When a query was submitted, the interface initiated a communication with the Okapi system through its Unix command line interface called *i1+*[2]. The i1+ performed the query processing such as stemming and stopword removal, weighting of query terms and, finally, returned a set of ranked documents back to the interface. The interface presented 10 documents per result page, and up to 200 documents were accessible per query.

For each retrieved document, the title and best passage were presented in the interface, along with the information on the document size, rank, and relevance score. The query terms were highlighted with bold in the title and best passage to increase an awareness of query terms' context in documents.

When a title was clicked, the full-text of the documents was shown (See Figure

[2]The manual of this interface can be found from `http://www.soi.city.ac.uk/~andym/OKAPI-PACK/` [Accessed: 6/1/07].

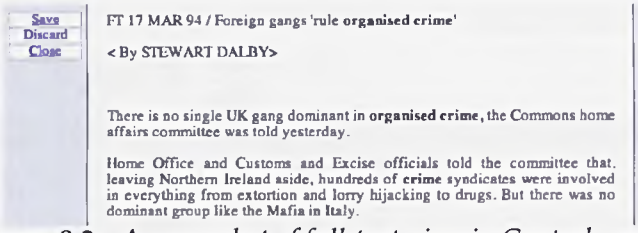



**Figure 8.3:** *A screenshot of full-text view in Control system*

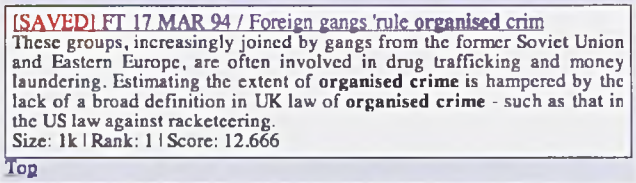



**Figure 8.4:** *Indication of the action taken on the document*

8.3). Participants were asked to take one of the three actions on the box shown next to the full-text: *Save*, *Discard*, and *Close*. Save was selected when a perceived relevant instance was found. Discard was selected when the document was perceived to be irrelevant. And, finally, Close was selected when neither of the previous two choices was appropriate. The saved and discarded documents were recorded, and an indication was made as shown in Figure 8.4 until a new query was submitted. The saved and discarded documents were removed from the result of subsequent queries. The indication was devised to facilitate browsing of search results.

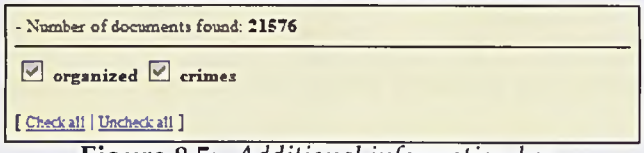



**Figure 8.5:** *Additional information box.*

The interface also showed a separate box (See Figure 8.5) above the first document to show the total number of documents that matched the query in the collection. This box also showed the current query terms to allow participants to modify the query by changing the checkbox next to each word. When a checkbox was unchecked, the corresponding word was removed from the query box at the top of the interface. This box was also used to present the history of the expanded query terms in the experimental systems, which will be discussed below.


**Figure 8.6:** *Experimental system: Initial screen.*

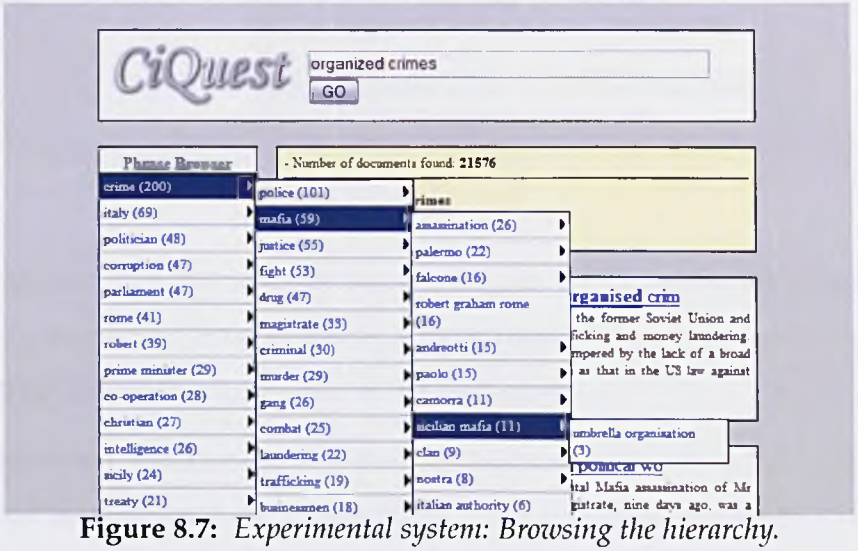



**Figure 8.7:** *Experimental system: Browsing the hierarchy.*

### 8.4.2 Experimental system

The experimental system was the same as the control system except the presence of the concept hierarchies and visual feedback of search terms. We first describe the difference between the Control system and Experimental system, followed by the description of the two methods used to generate the hierarchy.

When a query was submitted to the interface, the hierarchies were generated and presented at the left side of the interface (See Figure 8.6). In our experiment, the top 200 documents were used to generate a hierarchy. When the generation of the hierarchy was completed, the top level of the hierarchy was shown. The number in the brackets showed the frequency of occurrence of the term in the top

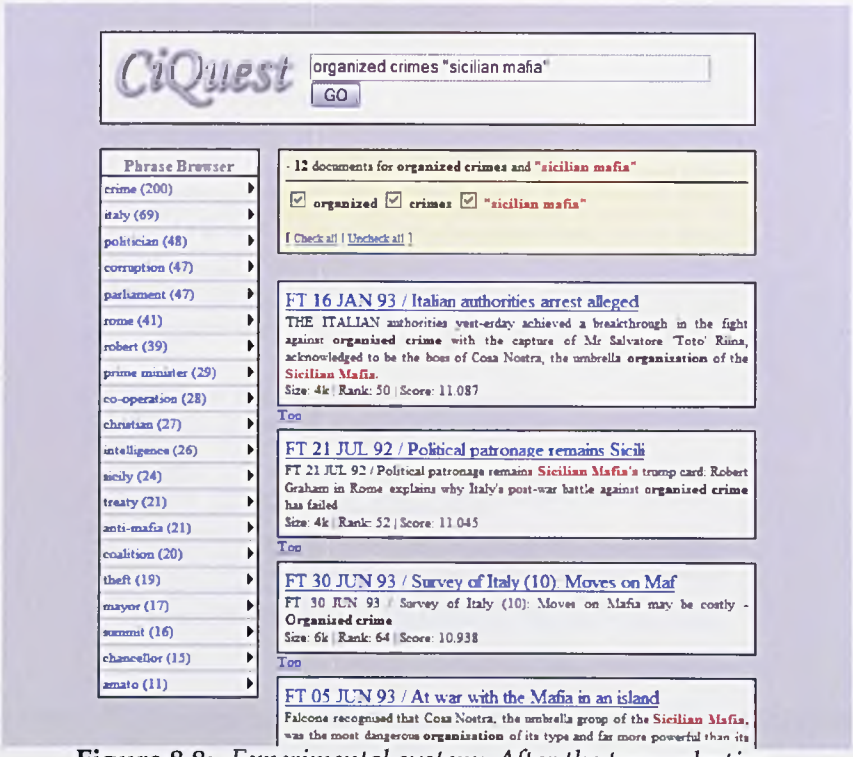



**Figure 8.8:** *Experimental system: After the term selection.*

200 documents. The frequency of occurrence was used to order the terms at each node of the hierarchy. In other words, a more frequent term was listed above than a lower term within the same node of the hierarchy.

The hierarchy was browsed by *hovering* the mouse pointer on the terms. When the pointer hovered on a term, a set of the child terms was shown in a sub-menu (See Figure 8.7). The presence of child terms was indicated by the rectangle arrow shown in the menu items.

When a term was *clicked* on the menu, a subset of retrieved documents that contained the selected term was shown (See Figure 8.8). The order of the subset documents was the same as the ranking of an initial retrieval. The selection of a term was seen as a form of query expansion in this study. Therefore, the selected terms are referred to *expansion terms* in the rest of this chapter. When a term was selected from the hierarchy, the information box was also updated with the expansion terms. The initial search result could be restored by clicking the "Show the original result" link at the bottom of the result page. When there were more than 10 documents for a selected term, the first 10 documents were presented in the same way as the initial result.

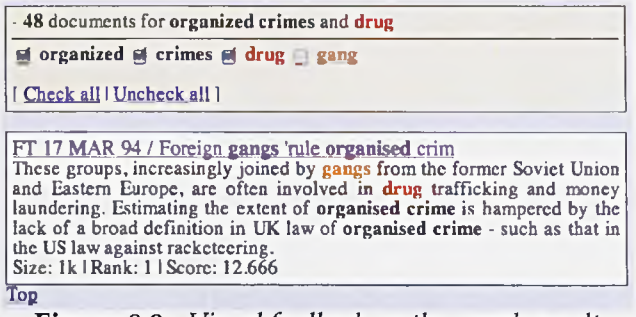



**Figure 8.9:** *Visual feedback on the search result*

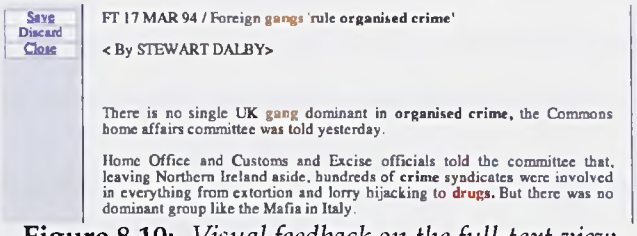



**Figure 8.10:** *Visual feedback on the full-text view*

To increase the awareness of the relationship between the initial query terms and expansion terms, a different visual feedback was devised for the experimental systems. In the control system, only the current query terms were highlighted in bold in the search results. In the experimental systems, however, the three types of highlighting were implemented as follows.

- **Black bold**: Current query terms (Control and Experimental)
- **Red bold**: Current expansion terms (Experimental)
- **Orange bold**: Previous expansion terms (Experimental)

The screenshots of the visual feedback are shown in Figure 8.9 and 8.10 for the search result and full-text view, respectively.

Compared to the control system, therefore, the experimental system was designed to increase the level of awareness regarding the relationship among the query terms, expansion terms, and retrieved documents in two ways: the presentation of terms in the hierarchy and varied visual feedback of terms in the search results. The relationship between the query terms and candidate expansion terms can be determined by the hierarchical menu, where the child terms can be seen as the related but more specific terms of the query terms. By looking at the terms at the same node, one can browse the terms which shared the same parent term. These

relationships were further contextualised in the search results using the different colours to highlight the terms.

In this study, we investigated the two underlying methods of the hierarchy generation based on a set of retrieved documents. The rest of this section describes the details of the hierarchy generation.

**Subsumption hierarchy** The subsumption hierarchy was a statistical approach to generating a concept hierarchy based on the co-occurrence information of terms. The generation of a subsumption hierarchy was implemented in the following way.

1. The full-text of the top 200 retrieved documents was fetched from the collection. We called this *retrieved set*.
2. Noun phrases were annotated in the texts of the retrieved set. The words and noun phrases were called the *terms*.
3. The document frequency and term frequency of the terms in the retrieved set were recorded. Stopwords and terms which appeared only once in the retrieved set were removed.
4. A document-term matrix was generated from the remained terms in the retrieved set.
5. For every pair of the terms, a parent-child relationship was determined by the subsumption conditions (the relaxed condition, See Section 4.3).
6. The parent-child pairs were combined to form a hierarchy.
7. For each node, the terms were ordered by the document frequency of the retrieved set. The top 20 terms was presented at each node.

In Step 2, the Brill tagger (Brill, 1992) was used to annotate the noun phrases in the retrieved set. The length of noun phrases was limited to three. In Step 3, the terms which occurred in only one document of the retrieved set was removed. The counting of frequency was duplicated for a noun phrase and the individual words of the phrase. In Step 6, some heuristics were used to resolve the pairs. For example, when there were three parent-child pairs such as $< A, B >$, $< A, C >$, and $< B, C >$, then $C$ was located in a sub-node of $B$ and only $B$ was located in a sub-node of $A$. Our implementation of the subsumption hierarchy was an extension of

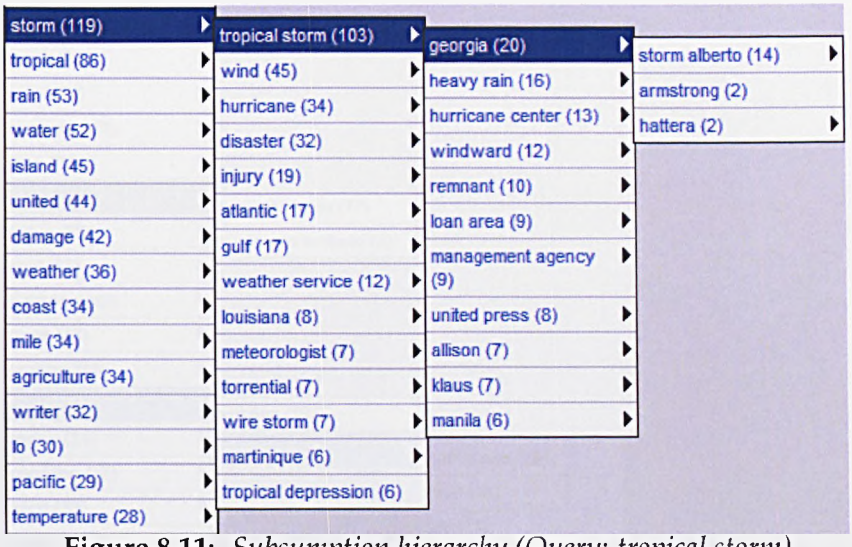



**Figure 8.11:** *Subsumption hierarchy (Query: tropical storm)*

the program developed in Sanderson and Croft (1999). Since the original program was designed to work in a batch mode process, some modification was applied to generate a hierarchy in a query time. For example, the size of the retrieved set was reduced from 500 to 200. The LCA terms extraction (Xu and Croft, 1996) were replaced by the Brill tagger's noun phrase extraction.

The example of the subsumption hierarchy is shown in Figure 8.11. The query was *tropical storm*. We can see the query terms occurred most frequently in the retrieved set. The noun phrase *tropical storm* was subsumed under *storm*, and we can find several names of tropical storms as a child of the phrase. The terms such as *rain*, *water*, and *island* appeared in the top level of the hierarchy.

**Keyphrase hierarchy** The keyphrase hierarchy was a lexical approach to generating a concept hierarchy. The keyphrase pattern matcher studied in Chapter 7 was used to extract parent-child pairs from the retrieved set. The generation of the keyphrase hierarchy was implemented in the following way.

1. The full-text of the top 200 retrieved documents was fetched from the collection (retrieved set).
2. Noun phrases were annotated in the texts of the retrieved set.
3. The document frequency and term frequency of the terms in the retrieved set were recorded.

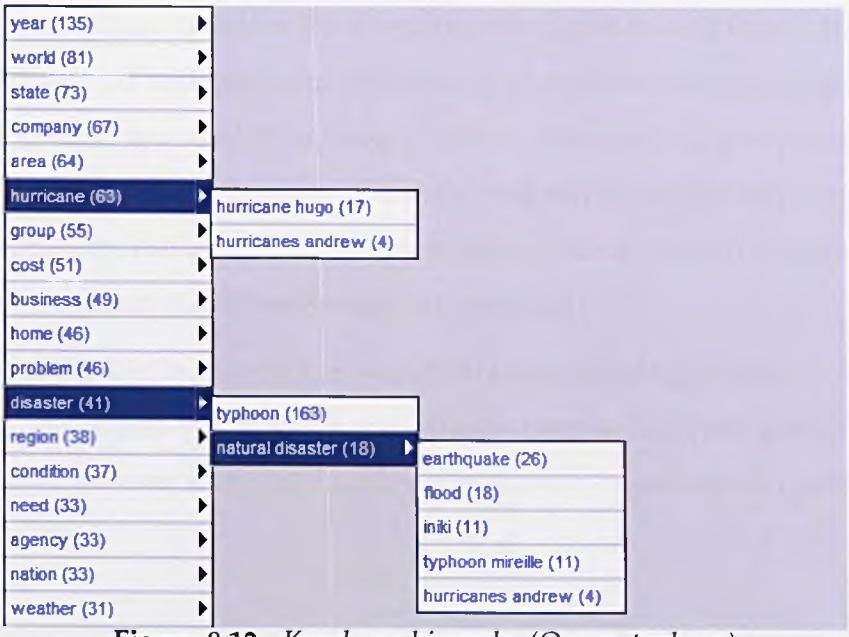



**Figure 8.12:** *Keyphrase hierarchy (Query: typhoon)*

4. The keyphrase pattern matching (See Section 7.2.1) was applied to the texts to extract the parent-child pairs.
5. The additional pairs were generated by a noun phrase and its head noun.
6. The parent-child pairs were combined to form a hierarchy.
7. For each node, the terms were ordered by the document frequency of the retrieved set. The top 20 terms was presented at each node.

Step 1 and 2 were the same as the subsumption hierarchy process. In Step 3, the stopwords removal and filtering were skipped because of the following pattern matching step. In Step 4, the pattern matching was applied without a predefined query noun. The annotation of the noun phrases was used to replace the descriptive phrase and query noun. In other words, when the sentence matched one of the patterns, the descriptive phrase was used as a parent of the query noun. For example, when a part of a matched sentence was "... natural disaster *such as* earthquake and flood ...", then the pairs $<$ natural disaster, earthquake $>$ and $<$ natural disaster, flood $>$ were extracted. In Step 5, when a phrase *natural disaster* was extracted, the head noun *disaster* was used to form the $<$ disaster, natural disaster $>$ pair. Step 6 and 7 were the same as the subsumption hierarchy process.

The example of the keyphrase hierarchy is shown in Figure 8.12. The query was *typhoon*. You can see the effect of the head noun pairs in *hurricane*. Compared

to the subsumption hierarchy, the keyphrase hierarchy tended to be smaller due to the frequency of matched texts. However, the semantic association between the parent and child appeared to be more intuitive. For example, the query term *typhoon* was located under the node of *disaster*, which was semantically correct. Due to this lexical-oriented approach, the parent terms did not necessarily have a higher document frequency like the subsumption hierarchy.

In summary, the two hierarchies were both automatically generated from a set of retrieved documents, but based on the different criteria. The next section presents the results of the user study which compared the two experimental systems to the control system.

## 8.5 Results and analysis

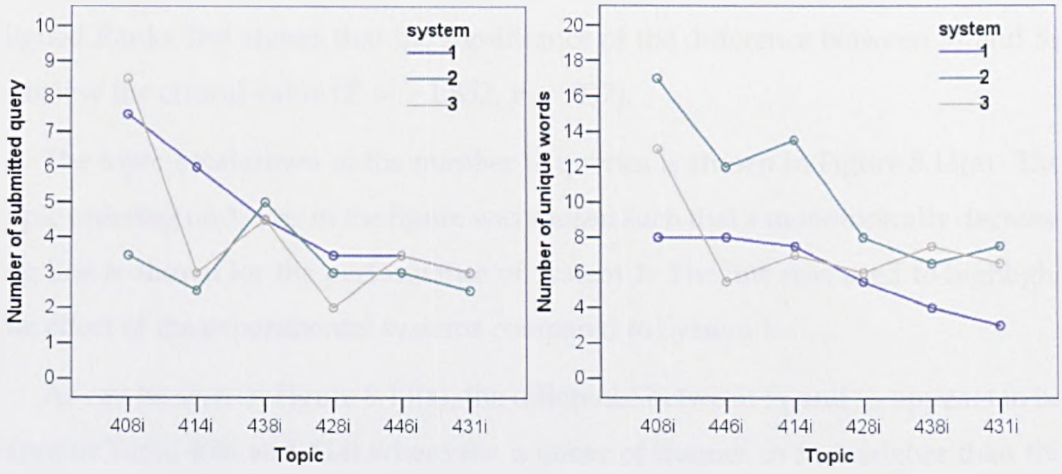
This section presents the experimental results of our study based on 36 search sessions carried out by 12 participants. The results presented in this section, unless otherwise stated, are the mean value of 12 sessions for one control and two experimental systems (denoted as $S_1$, $S_2$, and $S_3$, respectively). The standard deviation of the mean values is given in the brackets. The Friedman Test was run to establish the statistical significance ($p \leq .05$) of the differences observed among the three systems. When a difference was found to be significant, the post hoc test (Wilcoxon Signed Ranks Test) was carried out to find a significant pair(s) through the multiple pairwise comparisons of the three systems. To take an appropriate control of Type I errors, the significance level was set to $p \leq .0167$[3] in the post hoc tests, based on the Bonferroni correction (Siegel and Castellan, 1988).

In addition, the analysis of variance (ANOVA) test was run to see if there was an interaction effect between three independent variables: system, topic, and order. We only report the result of interaction effects when one of the independent variables is found to be significant and interaction effect is also found to be significant (i.e., $p \leq .05$). The ANOVA test was also used to establish the statistical significance ($p \leq .05$) of the differences observed between System 1 and the cumulate of System 2 and 3 (denoted as $S_{2+3}$). The motivation behind this was to isolate the effect of

[3]That is .05 divided by 3 pairwise comparisons.

**Table 8.2:** *Query reformulation*

| | System 1 | System 2 | System 3 | System 2+3 |
|---|---|---|---|---|
| Queries | 4.7 (2.1) | 3.3 (1.3) | 4.1 (2.3) | 3.7 (1.9) |
| Unique words | 6.0 (2.6) | 10.8 (5.6) | 7.6 (3.0) | **9.2** (4.7) |




(a) Queries (b) Unique words

**Figure 8.13:** *Topic-breakdown of query reformulation*

the hierarchies available in System 2 and 3 in the analysis. Finally, the Wilcoxon Signed Ranks Test was run when the effects of the hierarchy generation methods (i.e., between System 2 and 3) were compared.

This section is structured to address each of the specific research hypotheses discussed in Section 8.2. The results of additional analysis are also presented where appropriate to gain further insight into the effectiveness of the experimental systems.

### 8.5.1 Query reformulation

The first aspect to investigate was the effect of the concept hierarchies on the query reformulation. The assumptions were that the concept hierarchies can reduce the amount of effort required for manual query reformulation, and that the concept hierarchies can diversify the range of vocabulary used to complete the task. The results were as follows.

[*H*.1] *Concept hierarchies reduce the amount of effort required for manual query reformulation to complete a task.* The 2nd row of Table 8.2 shows the number of queries submitted to the systems during the tasks. As can be seen, the number of queries

appears to decrease in System 2 ($S_2$) and System 3 ($S_3$) compared to System 1 ($S_1$). The Friedman Test shows that the differences among the three systems are not statistically significant. And, the ANOVA shows that the difference between $S_1$ and the cumulate of System 2 and 3 ($S_{2+3}$) is not significant. However, the Wilcoxon Signed Ranks Test shows that the significance of the difference between $S_1$ and $S_2$ is below the critical value ($Z = -1.852$, $p = .037$).

The topic-breakdown of the number of queries is shown in Figure 8.13(a). The topic ordering on X-axis in the figure was chosen such that a monotonically decreasing line is shown for the performance of System 1. The line was used to highlight the effect of the experimental systems compared to System 1.

As can be seen in Figure 8.13(a), the difference between $S_1$ and $S_2$ appears to be large in Topic 408i and 414i where the number of queries in $S_1$ is higher than the other topics. This suggests that the effect of $S_2$ on the number of queries depends on the topics, but the effect is likely to increase when participants need to submit more queries to complete the task. While the trend of $S_3$ appears to be less consistent across the topics, there are more cases where the number of queries decreases than increases compared to $S_1$.

The ANOVA test shows that there is a significant interaction effect between the topic and its order ($F = 3.729$, $p = .017$). Further examination suggests that the number of queries is likely to vary across the presentation order on Topic 408i. More specifically, the number of queries in Topic 408i was noticeably smaller when it was presented last (Mean: 3.5) compared to when it was presented first (Mean: 7.5). The topic rotation shown in Table 8.1 informed us that the order of Topic 408i was always the same for the three systems. In other words, Topic 408i was tested first with System 1, tested second with System 3, and tested last with System 2 in our experiment. Therefore, while an interaction effect was found between the topic and its order, this might be a product of system effect indicated for Topic 408i. However, since other combinations of order were not tested for Topic 408i, the implication was not conclusive. We did not find an interaction effect between system and topic, and between system and order.

Overall, there appears to be some effect of the concept hierarchies on the number of manually reformulated queries. However, since an interaction effect was found

on one of the topics, and the difference among the systems was relatively small for four of the topics, [$H.1$] was not supported.

[$H.2$] *Concept hierarchies diversify the range of search vocabulary employed to complete a task.* The 3rd row of Table 8.2 shows the number of unique words used to complete the task. In $S_1$, this is the total number of unique words used in the queries. In $S_2$ and $S_3$, the number includes the words of the selected terms in the hierarchies during the task. As can be seen, the range of vocabulary appears to increase in $S_2$ and $S_3$ compared to $S_1$. The Friedman Test shows that the differences among the three systems are not significant. However, the ANOVA shows that the difference between $S_1$ and $S_{2+3}$ is significant ($F = 4.692$, $p = .037$). Therefore, the concept hierarchies had an effect of increasing the search vocabulary. Also, the Wilcoxon Signed Ranks Test shows that the significance of the difference between $S_1$ and $S_2$ is below the critical value ($Z = -2.451$, $p = .006$).

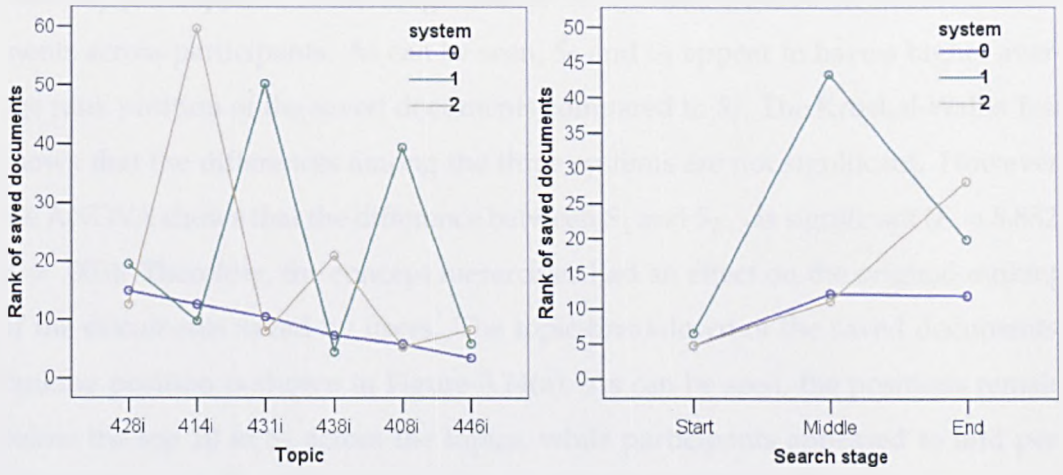
The topic-breakdown of the number of unique words is shown in Figure 8.13(b). As can be seen, the number of unique words in $S_2$ is consistently higher than $S_1$ across the topics. Similar to the number of queries, the difference appears to increase when participants need to use more unique words to complete the task in $S_1$ (i.e., Topic 408i, 446i, and 414i). The trend of $S_3$ appears to be different from $S_2$, but there are three topics where the number of unique words increases and one that decreases. These trends appear to support the significant difference between the systems discussed above.

The ANOVA test shows that there is a significant interaction effect between the topic and its order ($F = 3.320$, $p = .027$). Similar to the number of queries discussed above, the number of unique words appears to differ across the order of Topic 408i. When the topic was presented first and tested by System 1, the average number of unique words was 8. When it was presented second and tested by System 3, the average number was 13. When it was presented last and tested by System 2, the average number was 17. However, there was no significant interaction effect between system and topic, and between system and its order.

Since there was an interaction effect on one of the topics, an implication of the result might not be entirely conclusive. However, other results suggest that there appears to be an effect of concept hierarchies on the number of unique words em-

**Table 8.3:** *Browsing of search results*

| | System 1 | System 2 | System 3 | System 2+3 |
|---|---|---|---|---|
| Rank of saved docs | 9.2 (10.2) | 22.2 (40.4) | 16.7 (29.0) | **20.0** (35.6) |
| Next pages | 4.4 (3.6) | **1.8** (1.4) | **1.7** (2.1) | **1.7** (1.7) |




(a) Topic breakdown (b) Search session stage breakdown

**Figure 8.14:** *Breakdown of saved documents' rank*

ployed to complete a task, the effect is relatively consistent for the rest of topics. The trend plotted in Figure 8.13(b) also suggests that the effect is likely to be consistent in System 2 compared to System 1. Therefore, we judged that [*H*.2] was supported.

It should be noted that an interaction effect between the topic and its order was not found in the rest of our results.

### 8.5.2 Browsing of retrieved documents

The second aspect of our investigation was the effect of the concept hierarchies on the browsing of retrieved documents. The assumptions were that the concept hierarchies assist a searcher in browsing the retrieved documents effectively, and that the concept hierarchies can increase the probability of finding perceived relevant information in click-through documents. The results are as follows.

[*H*.3] *Concept hierarchies encourage participants to view the documents that are retrieved in a low rank.* The concept hierarchies in $S_2$ and $S_3$ were designed to support an effective browsing of the retrieved documents by grouping the search results based on the terms in the hierarchy. The grouping of the retrieved documents tended to bring a lower ranked document up in the results. Therefore, by looking

at the average rank of saved documents, one can examine the effect of the concept hierarchies on the effective browsing of the retrieved documents.

The second row of Table 8.3 shows the average original rank of the documents saved by participants. The average rank of the systems is based on all saved documents across participants. As can be seen, $S_2$ and $S_3$ appear to have a higher average rank position of the saved documents compared to $S_1$. The Kruskal-Wallis Test shows that the differences among the three systems are not significant. However, the ANOVA shows that the difference between $S_1$ and $S_{2+3}$ is significant ($F = 8.882$, $p = .003$). Therefore, the concept hierarchies had an effect on the original ranking of the documents saved by users. The topic-breakdown of the saved documents' ranking position is shown in Figure 8.14(a). As can be seen, the positions remain below the top 20 in $S_1$ across the topics, while participants appeared to find perceived relevant information in a deeper position in $S_2$ and $S_3$. The effect appears to vary across the topics in the experimental systems.

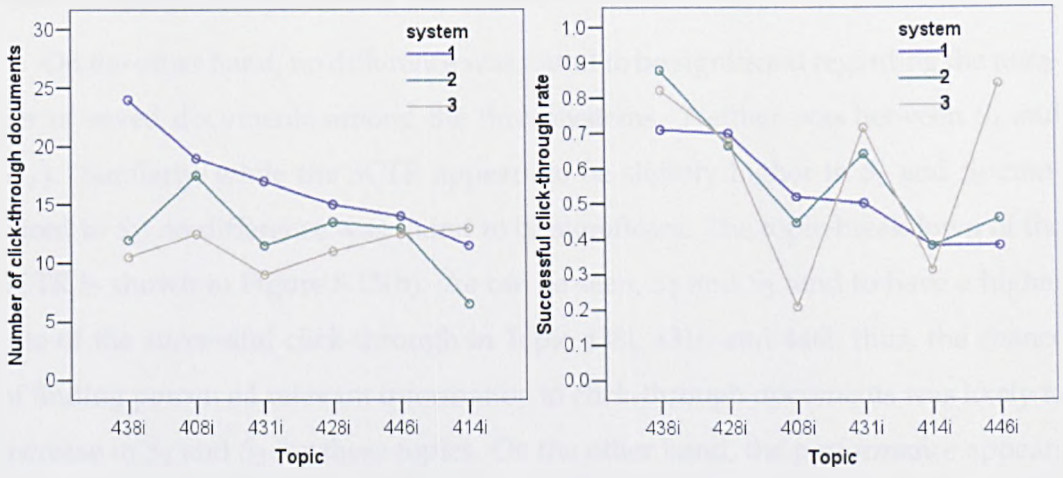
This effect was further supported by the frequency of viewing the *next* result pages (i.e., after the first result page) showing in the bottom row of Table 8.3. The Friedman Test shows that the differences among the three systems are significant ($\chi^2(2) = 8.227$, $p = .014$). The posthoc tests show that the difference is significant between $S_1$ and $S_2$ ($Z = -2.585$, $p = .004$) and between $S_1$ and $S_3$ ($Z = -2.708$, $p = .002$). The ANOVA shows that the difference between $S_1$ and $S_{2+3}$ is significant ($F = 9.417$, $p = .004$). Therefore, participants viewed a significantly fewer number of the extra result pages in $S_2$ and $S_3$ compared to $S_1$.

Figure 8.14(b) plots the mean rank position of the saved documents across the stage of a search session. The stage was determined by dividing the entire search session into the three regions of equal length: start, middle, and end. For example, when a subject took nine minutes to complete a task, then each stage consisted of three minutes long. As can be seen, the position of the saved documents in $S_1$ remains mostly below a certain level while the positions in $S_2$ and $S_3$ appear to vary more widely. In particular, the middle stage of a search session tended to have a higher rank of saved documents in $S_2$.

Overall, these results show that the concept hierarchies helped participants find relevant information in a lower ranked document without viewing the extra search

**Table 8.4:** *Successful click-through rate*

| | System 1 | System 2 | System 3 | System 2+3 |
|---|---|---|---|---|
| Click-through | **16.8** (5.5) | 12.3 (3.9) | 11.4 (3.4) | 11.9 (3.6) |
| Saved documents | 9.1 (4.7) | 7.2 (3.1) | 6.2 (3.2) | 6.7 (3.1) |
| Successful click-through | .53 (.18) | .58 (.19) | .60 (.31) | .59 (.25) |




(a) Click-through (b) Successful click-through rate

**Figure 8.15:** *Topic breakdown of Click-through and its saved portion*

result pages, thus, [*H*.3] was supported. The result also shows that the effect of the concept hierarchies on the browsing of search results can vary across the stage of a search session.

[*H*.4] *Concept hierarchies improve the successful click-through rate.* [*H*.3] examined browsing of retrieved documents based on the original rank position of saved documents. Another way to investigate the effect of the concept hierarchies on the browsing of retrieved documents was the successful click-through rate (SCTR), which was a portion of saved documents in click-through documents. The assumption was that an effective browsing of search results should increase the probability of viewing relevant documents.

The second to fourth rows of Table 8.4 show the number of click-through documents, number of saved documents, and successful click-through rate, respectively. As can be seen, the number of click-through documents appears to be higher in $S_1$ than $S_2$ and $S_3$. The Friedman Test shows that the differences among the three systems are significant ($\chi^2(2) = 8.227$, $p = .014$), and the posthoc tests show that the difference is significant between $S_1$ and $S_2$ ($Z = -2.585$, $p = .004$) and between $S_1$

and $S_3$ ($Z = -2.708$, $p = .002$). The ANOVA shows the same result ($F = 10.150$, $p = .003$). Therefore, participants viewed more documents in $S_1$ compared to $S_2$ and $S_3$. Figure 8.15(a) shows that the number of click-through in $S_1$ is consistently higher than the other two systems across the topics. The difference between the systems appears to be larger in Topic 438i and 431i.

On the other hand, no difference was found to be significant regarding the number of saved documents among the three systems. Neither was between $S_1$ and $S_{2+3}$. Similarly, while the SCTR appears to be slightly higher in $S_2$ and $S_3$ compared to $S_1$, no difference was found to be significant. The topic-breakdown of the SCTR is shown in Figure 8.15(b). As can be seen, $S_2$ and $S_3$ tend to have a higher rate of the successful click-through in Topic 438i, 431i, and 446i, thus, the chance of finding perceived relevant information in click-through documents was likely to increase in $S_2$ and $S_3$ for these topics. On the other hand, the performance appears to be degraded in Topic 408i. Therefore, the small difference of the SCTR among the systems appears to be the product of the varied performance across the topics. In summary, our conservative decision was that $[H.4]$ was not supported.

The ANOVA test shows that there is a significant interaction effect between system and topic on the number of saved documents ($F = 2.812$, $p = .027$). Further examination suggests that the number of saved documents in System 3 is noticeably lower than the other two systems on Topic 408i. Therefore, participants might have found System 3 difficult to find relevant information on Topic 408i. An indication is also found in the topic-breakdown of the SCTR shown in 8.15(b).

### 8.5.3 Task performance and perception

The results so far have suggested that, with the concept hierarchies, participants tended to employ an increased number of words to complete the task compared to the control system ($S_1$). The results have also suggested that the browsing of retrieved documents was facilitated by the concept hierarchies with a varied degree across the topics. This section investigates if these effects of the concept hierarchies contribute to the performance of the instance finding task. The assumption was that the concept hierarchies had a positive effect both on the task performance and

participants' perceptions on the task. Since the hierarchy was designed to allow participants to focus on a particular set of retrieved documents based on the terms in a hierarchy, the precision was expected to improve. The results are as follows.

[*H*.5] *Concept hierarchies improve the task performance, especially in precision.* The instance finding task asked participants to find as many different instances that were relevant to a topic as possible. The instance recall was defined by the portion of correctly identified instances in the official instances identified by the NIST assessors. The instance precision, on the other hand, was defined by the portion of correct instances in the retrieved instances. We also measured the E-score (Van Rijsbergen, 1979) which was a single score based on recall and precision. The E-score is defined as follows.

$$E = 1 - \frac{1}{\alpha \frac{1}{Precision} + (1-\alpha)\frac{1}{Recall}} \tag{8.5}$$

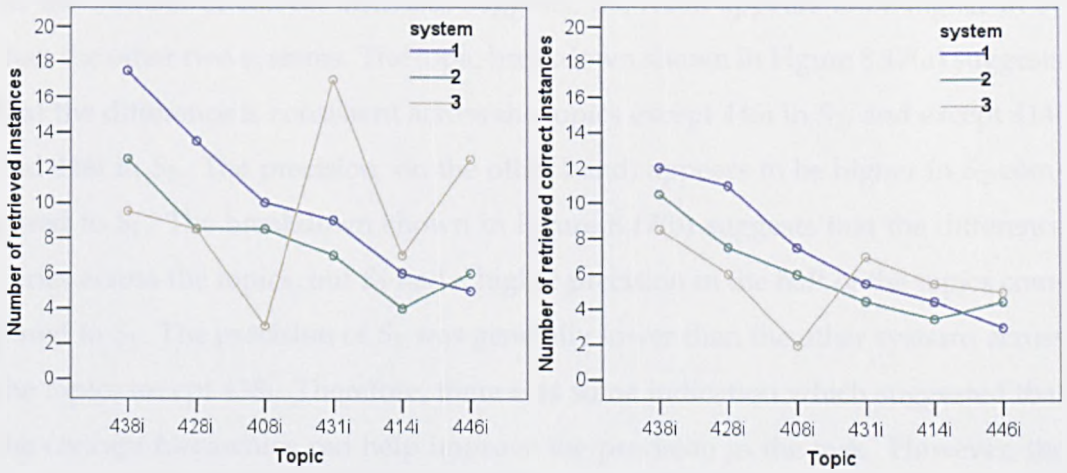
The E-score allowed us to vary the weight of recall and precision by changing the $\alpha$ parameter. With $\alpha = .5$, the recall and precision are given the same weight. With $\alpha = .2$, the E-score weights the recall twice as much as precision. With $\alpha = .8$, the score weights the precision twice as much as recall. This simulates a recall-oriented or precision-oriented search in the performance measure. Since the E-score is a distance measure, a smaller value represents a better performance between 0 and 1. To be comparable with recall and precision representations, we show the score $1 - E$ in the following result.

In addition, we measure the time taken to complete the task for the task performance. The results are shown in Table 8.5.

The second and third rows of Table 8.5 show the number of retrieved instances and that of correct instances. As can be seen, there is a comparable number of instances retrieved by $S_1$ and $S_3$, and $S_2$ appears to be somewhat lower. On the other hand, the number of retrieved correct instances is similar in $S_2$ and $S_3$, but both appear to be lower than $S_1$. However, no difference was found to be significant. The only exception was that Wilcoxon Signed Ranks Test showed that the difference between $S_1$ and $S_2$ was below a critical value regarding the number of retrieved instances ($Z = -1.793$, $p = .045$). Therefore, participants were likely to retrieve

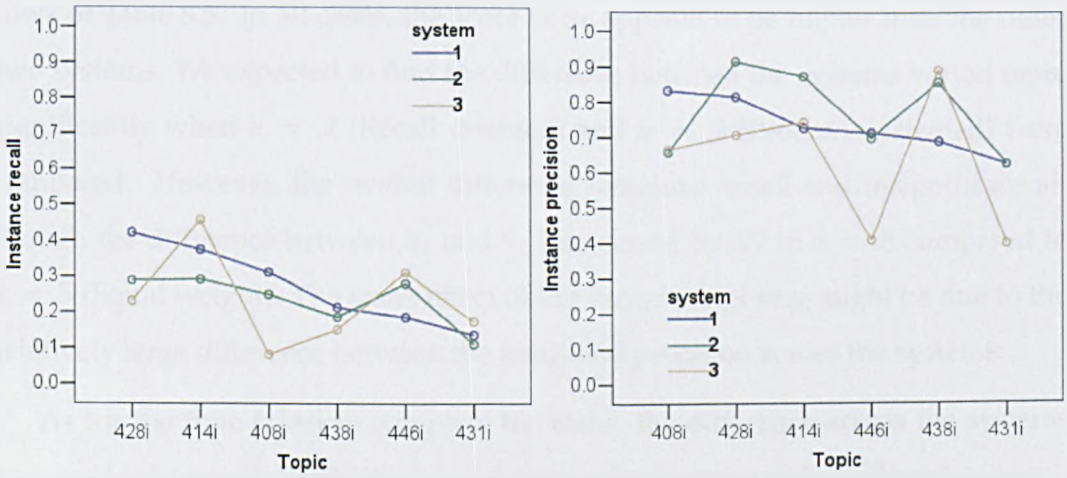
**Table 8.5:** *Task performance*

| | System 1 | System 2 | System 3 | System 2+3 |
|---|---|---|---|---|
| Retrieved instances | 10.2 (5.2) | 7.8 (3.4) | 9.6 (6.2) | 8.7 (5.0) |
| Correct instances | 7.3 (3.8) | 6.1 (3.2) | 5.7 (3.0) | 5.9 (3.1) |
| Instance Recall | .28 (.13) | .24 (.12) | .24 (.16) | .24 (.14) |
| Instance Precision | .73 (.18) | .77 (.21) | .64 (.24) | .70 (.23) |
| $1 - E$ score ($\alpha = .5$) | .39 (.14) | .35 (.15) | .31 (.17) | .33 (.16) |
| $1 - E$ score ($\alpha = .2$) | .31 (.14) | .27 (.13) | .26 (.17) | .27 (.15) |
| $1 - E$ score ($\alpha = .8$) | .53 (.14) | .51 (.18) | .43 (.17) | .47 (.18) |
| Completion time (sec) | 599.6 (1.0) | 593.9 (7.8) | 598.3 (3.2) | 596.1 (6.2) |




(a) Retrieved instances (b) Retrieved correct instances

**Figure 8.16:** *Topic breakdown of retrieved instances*




(a) Instance recall (b) Instance precision

**Figure 8.17:** *Topic breakdown of instance recall and precision*

more instances in $S_1$, but there was no significant difference between $S_1$ and $S_2$ in the number of correct instances.

The topic-breakdown of the results is shown in Figure 8.16. The breakdown shows that $S_1$ and $S_2$ have a similar trend across the topics except Topic 446i. The performance of $S_3$ in the number of retrieved documents appears to vary differently. The peak on Topic 431i of $S_3$ in Figure 8.16(a) is due to a participant who retrieved 26 instances (of which 11 were correct). As for the number of retrieved correct instances, again, the trend of $S_1$ and $S_2$ appears to be similar across the topics except 446i. $S_3$ appears to retrieve more correct instances than $S_1$ in three topics.

The fourth and fifth rows of Table 8.5 show the true instance recall and precision. As the number of correct instances suggests, the recall appears to be higher in $S_1$ than the other two systems. The topic-breakdown shown in Figure 8.17(a) suggests that the difference is consistent across the topics except 446i in $S_2$, and except 414i and 446i in $S_3$. The precision, on the other hand, appears to be higher in $S_2$ compared to $S_1$. The breakdown shown in Figure 8.17(b) suggests that the difference varies across the topics, but $S_2$ had a higher precision in the half of the topics compared to $S_1$. The precision of $S_3$ was generally lower than the other systems across the topics except 438i. Therefore, there was some indication which suggested that the concept hierarchies can help improve the precision in the task. However, the overall difference tended to be small, and no difference was found to be significant.

The $1 - E$ scores with the varied weighting are shown in the sixth to eighth rows of Table 8.5. In all cases, the score of $S_1$ appears to be higher than the other two systems. We expected to find the difference between the systems varied more significantly when $\alpha = .2$ (Recall oriented) and $\alpha = .8$ (Precision oriented) were compared. However, the overall difference remained small and insignificant although the difference between $S_1$ and $S_2$ was closed by .02 in $\alpha = .8$ compared to $\alpha = .5$ (Equal weight). The small effect of the varied weighting might be due to the relatively large difference between the recall and precision across the systems.

As for the time taken to complete the tasks, the difference across the systems appeared to be small. Participants used 10 minutes for most tasks in the experiment. Participants on average appeared to complete the tasks in a slightly shorter time in $S_2$ compared to $S_1$ and $S_3$. And, the Friedman Test shows that the difference among the systems is significant ($\chi^2(2) = 9.852$, $p = .004$). However, the practical effect of this difference should be negligible.

**Table 8.6:** *Task perceptions*

| | System 1 | System 2 | System 3 | System 2+3 |
|---|---|---|---|---|
| Ease of getting started | 5.9 (0.9) | 5.8 (0.9) | 5.3 (1.2) | 5.5 (1.1) |
| Ease of overall search | 5.3 (1.2) | 5.2 (1.3) | 4.8 (1.1) | 5.0 (1.2) |
| Outcome satisfaction | 5.3 (1.1) | 4.9 (1.2) | 4.1 (1.7) | 4.5 (1.5) |
| Search time | 5.3 (1.2) | 5.1 (1.5) | 4.4 (1.7) | 4.9 (1.5) |

In summary, the task performance of the systems appears to be comparable, thus, [$H$.5] was not supported.

[$H$.6] *Concept hierarchies enable users to perceive that the searches are easy and successful.* Table 8.6 shows the results of participants' perceptions on the tasks they carried out. A 7-point scale was used to capture the subjective assessments on the tasks where a higher score represented a better assessment in the analysis. For example, the 2nd row of the table (Ease of getting started) is the result of the question: *Was it easy to get started on this search?*, and the scale was 1 (Not at all) to 4 (Somewhat) to 7 (Extremely). Similarly, the questions were: *Was it easy to do the search on this topic?* (Ease of overall search), *Are you satisfied with your search results?* (Outcome satisfaction), *Did you have enough time to do an effective search?* (Search time) for the 3rd, 4th, and 5th row of the table, respectively.

As can be seen, participants appeared to indicate a better assessment on $S_1$ across the questions although the results of $S_1$ and $S_2$ often appeared to be comparable. Participants' assessments on the tasks with $S_3$ tended to be lower than the other two systems. However, no difference was found to be significant in all results. The difference of the ease of getting started had the lowest $p$ value ($p = .26$) based on the Friedman Test, and others varied from $p = .48$ (Search time) to $p = .71$ (Ease of overall search). The ANOVA shows the same tendency where the $p$ value ranged from $p = .10$ (Outcome Satisfaction) to $p = .44$ (Ease of overall search). Therefore, we did not find a significant effect of the concept hierarchies on participants' task perceptions, thus, [$H$.6] was not supported.

### 8.5.4 Information seeking behaviour

The results shown so far have focused on the effect of the concept hierarchies compared to the control system. This section investigates participants' information

**Table 8.7:** *Hierarchy access during a search session.*

| Search stage | System 2 | System 3 | System 2+3 |
|---|---|---|---|
| Start | 15 | 15 | 30 |
| Middle | 17 | 6 | 23 |
| End | 22 | 9 | 31 |
| Total | 54 | 30 | 84 |

seeking behaviour with the concept hierarchies. In particular, the frequency of the hierarchy use over the stage of a search session, the depth of selected terms in the hierarchy, and task perceptions that affect the use of the concept hierarchies are examined. The results are as follows.

[*H.7*] *The frequency of the access to the hierarchies changes over the stage of a search session.* As the search progresses, it is likely that a searcher learns more about the search topic and retrieved documents. At the same time, however, it is also likely that the number of unseen relevant instances in the collection decreases. This makes the search more difficult than the previous stage, thus, users might seek support to find relevant information. Table 8.7 shows the accumulated number of accesses to the hierarchies over the three stages of a search session. The stages were determined in the same manner as Figure 8.14(b) (the method is given in page 126).

As can be seen, the frequency of hierarchy access increases in $S_2$ towards the end of search. On the other hand, the beginning of search was the most frequent stage of access in $S_3$. The table also shows that the hierarchies in $S_3$ were not as frequently used as $S_2$ from the middle to end stage of search. The Spearman's correlation coefficient shows that the frequency of hierarchy access are significantly correlated with the stage of a search session in $S_2$ ($p = .000$), but not $S_3$ or $S_{2+3}$. Therefore, [*H.7*] was partially supported.

The average frequency of the hierarchy access based on the overall search sessions was 4.5 (SD: 3.1) and 2.5 (SD: 1.8) for $S_2$ and $S_3$, respectively. The Wilcoxon Signed Ranks Test shows that the difference between the two systems is significant ($Z = -2.012$, $p = .047$). Therefore, participants used the hierarchies in $S_2$ more frequently than $S_3$.

[*H.8*] *The depth of selected concepts in the hierarchies changes over the stage of a search session.* For a similar reason to the previous hypothesis, we were also interested in

**Table 8.8:** *The depth of selected terms in the hierarchy.*

| Search stage | System 2 | System 3 | System 2+3 |
|---|---|---|---|
| Start | 3.6 (1.1) | 2.8 (1.3) | 3.3 (1.2) |
| Middle | 3.2 (1.0) | 2.3 (1.4) | 3.0 (1.1) |
| End | 3.6 (0.7) | 2.8 (1.0) | 3.2 (1.0) |
| Total | 3.5 (1.0) | 2.7 (1.1) | 3.2 (1.1) |

**Table 8.9:** *Correlation between the hierarchy use and task perceptions.*

| | Ease of start | Ease of search | Previous knowledge |
|---|---|---|---|
| Hierarchy use | -.027 | -.271 | **-.313** |

the relationship between the depth of the selected terms in the hierarchy and stage of a search session. An increased level of knowledge on the topic and decrease of relevant information in the collection might affect the depth of the hierarchies accessed by a searcher. Table 8.8 shows the average depth of the selected terms in the hierarchy over the stage of a search session. As can be seen, contrary to our assumption, the depth rarely varied over the stage. The total standard deviation was also found to be relatively low. The table also shows that the most frequently accessed levels were 3 and 4 in $S_2$ and 2 and 3 in $S_3$. Therefore, [*H*.8] was not supported.

[*H*.9] *The frequency of the access to the hierarchies correlates with a searcher's perceptions of search topics.* The access to the hierarchies can be seen as an indication of a searcher's need for help in the search. Therefore, we measured the correlation between the frequency of the access to the hierarchies and task perceptions which have been shown in Table 8.6. In addition, we included participants' subjective assessment on the previous knowledge of the task topics. The assessment was captured in the same fashion as the other task perceptions. The question was *Did your previous knowledge help you with your search?* This assessment was used as an indication of the level of familiarity with the topics.

The result of Spearman's correlation co-efficient (2-tailed) is shown in Table 8.9. As can be seen, the ease of getting started, ease of overall search, and previous knowledge appeared to negatively correlate with the frequency of the hierarchy use. This suggests that participants were likely to access the hierarchy when the search was perceived to be more difficult. The negative correlation with the level of previous knowledge is shown to be significant ($p = .045$). Therefore, [*H*.9] was

partially supported.

To gain further insight into the intention behind the use of the hierarchy, we carried out an exploratory analysis on the interaction path extracted from the search logs. The analysis of interaction path has been used, for example, to develop a relevancy prediction model from users' interactions (Fox et al., 2005). While there can be a different way to analyse the interaction path, we used a simple method which looked at the subsequent search actions taken during the experiment. We extracted eight representative actions from our search logs. For each action, we counted the number of cases where the other actions followed. Table 8.10 shows the results of analysis.

The paths had a starting point (denoted as *START*) and end point (*END*). The number next to the action code is the frequency of occurrence. The frequency was based on the accumulation of all interactions recorded for each system. For example, in $S_1$, there were a total of 50 queries submitted to the system (denoted as *Search*). The next actions of *search* consisted of the viewing a full-text (*view*), resubmitting a new query, or viewing the next search result page (*next*). The table shows that 74% of cases, participants viewed a full-text after a new query.

There are some trends of the hierarchy use that have emerged from the analysis of the interaction path. Firstly, the access to the hierarchy tended to replace the manual reformulation of queries. In $S_1$, the *Search* action was followed by another *search* action in 16% of cases. This portion decreases to 8% in $S_2$ and the probability of the *expand* action after the *Search* action was 19%. While we did not find a significant effect of concept hierarchies on reducing the number of manual query reformulation, this appears to suggest that participant accessed the hierarchy rather than submitting a new query. Secondly, the access to the hierarchy tended to replace the next page viewing. For example, the *Save* and *Close* actions in $S_1$ had the *next* (page) action in 18% and 22% of cases, respectively. However, in $S_2$, both actions had the *expand* action in 16% and 30% of cases. This appears to support the result of an decreased number of the next page viewing discussed in Section 8.5.2. Finally, an unsuccessful click-through tended to motivate the use of the concept hierarchies. For example, in $S_2$ and $S_3$, the probability of the *expand* action after the *Close* was twice as high as the *Save* action. This appears to support the result of the task per-

**Table 8.10:** *Information seeking path: accumulated frequency of next actions.*

| START ($S_1$) | 12 | Search | 50 | View | 201 | Save | 109 | Close | 77 | Discard | 12 | Next | 53 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| search | 1.0 | view | .74 | save | .54 | view | .77 | view | .56 | search | .42 | view | .55 | | | | |
| | | search | .16 | close | .38 | next | .18 | next | .22 | view | .42 | next | .21 | | | | |
| | | next | .10 | discard | .06 | search | .05 | search | .14 | END | .17 | search | .17 | | | | |
| | | | | †other | .01 | | | END | .08 | | | END | .08 | | | | |
| START ($S_2$) | 12 | Search | 37 | View | 148 | Save | 86 | Close | 56 | Discard | 6 | Next | 16 | Expand | 54 | Ex-nxt | 5 |
| search | 1.0 | view | .70 | save | .58 | view | .57 | view | .43 | view | .33 | view | .69 | view | .59 | view | .80 |
| | | expand | .19 | close | .38 | expand | .16 | expand | .30 | next | .33 | search | .19 | expand | .22 | expand | .20 |
| | | search | .08 | discard | .04 | next | .12 | END | .11 | search | .17 | expand | .13 | search | .15 | | |
| | | END | .03 | | | search | .07 | search | .07 | expand | .17 | | | next | .02 | | |
| | | | | | | ex-nxt | .03 | next | .05 | | | | | END | .02 | | |
| | | | | | | END | .05 | ex-nxt | .04 | | | | | | | | |
| START ($S_3$) | 12 | Search | 47 | View | 137 | Save | 74 | Close | 60 | Discard | 3 | Next | 14 | Expand | 30 | Ex-nxt | 6 |
| search | 1.0 | view | .66 | save | .54 | view | .61 | view | .52 | view | .67 | view | .50 | view | .53 | view | .83 |
| | | search | .17 | close | .44 | next | .12 | search | .23 | expand | .33 | next | .21 | expand | .20 | END | .17 |
| | | expand | .15 | discard | .02 | expand | .08 | expand | .17 | | | search | .14 | search | .20 | | |
| | | END | .02 | | | search | .07 | next | .03 | | | END | .14 | END | .07 | | |
| | | | | | | END | .07 | ex-nxt | .03 | | | | | | | | |
| | | | | | | ex-nxt | .05 | END | .02 | | | | | | | | |

*START*: Beginning of task. *Search*: New query. *View*: Full-text access. *Save/Close/Discard*: Relevance judgement. *Next*: Viewing the next result page. *Expand*: Term selection in a hierarchy. *Ex-nxt*: Viewing the next page on the expanded results. *END*: End of task.

† Full-text window closed without a judgement.

**Table 8.11:** *Participants' perceptions on the hierarchies.*

| | | System 2 | System 3 | Wilcoxon | T-test |
|---|---|---|---|---|---|
| Q1. | Was it easy to browse the hierarchical menu? | 4.8 (1.0) | 4.3 (1.7) | .344 | .244 |
| Q2. | Was the menu too deep/long to manage? | 3.3 (1.7) | 2.9 (1.5) | .672 | .479 |
| Q3. | Was the menu confusing or misleading? | 2.1 (1.0) | 3.7 (1.4) | .391 | .280 |
| Q4. | Were the terms in the menu helpful for predicting the contents of the linked documents? | 4.8 (0.8) | 4.2 (1.7) | .375 | .370 |
| Q5. | Were the terms in the menu helpful for judging the relevance of documents? | 4.5 (0.9) | 4.3 (1.8) | .781 | .693 |
| Q6. | Was the menu helpful in focusing on important terms? | 5.0 (1.6) | 4.2 (1.8) | .230 | .182 |
| Q7. | Were the terms in the menu helpful for understanding the retrieved documents? | 4.0 (1.6) | 2.8 (1.1) | **.031** | **.013** |
| Q8. | By browsing the menu, did you feel that you had a better idea of the contents of a set of retrieved documents? | 4.7 (1.3) | 3.7 (1.6) | **.047** | **.023** |

ceptions' effect on the use of the hierarchies, which was discussed in Section 8.5.4.

One of the observations made by the experimenter during the study was that participants appeared to browse the terms at the same node frequently. Therefore, we also analysed the logs to examine the observation. There was a total of 67 term selections made in the hierarchies that were deeper than the second degree (i.e., Top > Level 1 > Level 2). Of those, 11 (16.4%) had more than one selection from the same node. Many of them were associated with the viewing and saving actions. This excludes the non-detectable browsing of the hierarchies. Therefore, a successful focus on a set of retrieved documents appeared to motivate participants to revisit the other members of the same node.

### 8.5.5 Participants' perceptions on the concept hierarchies

The previous analysis helped us understand the factors that affected the use of concept hierarchies during the task. This section presents the results of participants' perceptions on the usability and usefulness of concept hierarchies to compare $S_2$ and $S_3$. The subjective assessments on the concept hierarchies were captured in the same fashion as before based on a 7-point scale. The significance of the difference was tested by the Wilcoxon Signed Ranks Test (2-tailed) and Paired T-test. Two participants did not complete the questions on the assessments of the hierarchies due to the little use of the hierarchies during the task. Since we were interested in a comparison of the experimental systems in this section, we decided to remove the two cases in the analysis. Therefore, the results are the mean of 10 searches. Table 8.11 shows the results of analysis. The first three questions shown in the second to fourth rows of the table were concerned with the usability of the hierarchies, while the rest of the questions were concerned with the usefulness of the hierarchies in the support of searches. The last two columns show the significance level of the difference determined by the tests.

The usability questions suggest that participants appear to find $S_2$ easier to browse the hierarchy (Q1) and less confusing or misleading (Q3) than $S_3$. However, it appears that participants found $S_2$'s hierarchy sometimes too deep or long to manage (Q2) compared to $S_3$. The result of Q2 was somehow expected since the subsumption hierarchies in $S_2$ tended to present more terms than $S_3$. However, it appears that the structure of the keyphrase hierarchy in $S_3$ was sometimes more confusing to interpret than $S_2$. This was unexpected since we assumed that the relation between the terms in the keyphrase hierarchies was more intuitive than the subsumption hierarchies. One of participants expressed a negative impression on the keyphrase hierarchy by pointing out an irrelevant term shown at the top level of the menu. This suggests that the terms shown at the top level can influence participants' perception on the hierarchies since only the top level was presented in the initial search results. Therefore, the selection of the top level terms can be important for the usability of the concept hierarchies. Due to the lexical oriented method of the keyphrase hierarchies, the query terms were sometimes located at a lower level of the menu (See Figure 8.12 in Section 8.4.2). This might be one of the

factors that confused participants in $S_3$. However, overall, no difference was found to be significant by both tests. Therefore, the usability of the two hierarchies was comparable.

Q4 to Q8 were the questions regarding the usefulness of the hierarchies. As can be seen, participants found $S_2$ more useful than $S_3$ across the questions. A significant difference was found in Q7 and Q8 by both tests. Q7 was about the usefulness of the hierarchies for understanding of retrieved documents, while Q8 was about the usefulness of the hierarchies for getting an overview of retrieved documents. The result shows that the subsumption hierarchies offered more terms which helped participants understand the content of retrieved documents compared to the keyphrase hierarchies. The result also shows that the subsumption hierarchies were more useful for getting an overview of the retrieved documents. It should be mentioned that the difference between the two questions appeared to be confusing to a couple of participants since they asked for a clarification. For these two questions, we run the ANOVA test to see if there was an interaction effect between system, topic, and order. No significant interaction effect was found.

The overall result suggests that participants tended to find the subsumption hierarchies more useful for completing the task than the keyphrase hierarchies. This appears to correlate with the frequency of the hierarchy use discussed in Section 8.5.4. The frequency of the hierarchy access in $S_3$ was lowered in the middle to end of search session, while the frequency increased towards the end of a search session in $S_2$.

#### 8.5.5.1 System preference and other feedback

At the end of three tasks, participants were asked to give their preference and other feedback on the control and experimental systems. The first question asked *How different did you find the systems from one another?* Participants' assessment was captured by a 5-point scale where 1 represented "not at all", 3 represented "somewhat", and 5 represented "extremely". The mean value of the assessment based on twelve participants was 2.6 (SD: 1.2). Of twelve, six gave Score 1 or 2 (three each), two gave 3, and finally, four gave 4. None gave Score 5. The Chi-square test shows that the difference is not significant. Therefore, participants appeared to find the

Table 8.12: *Post-test questions.*

| | N | System 1 | System 2 | System 3 | No difference |
|---|---|---|---|---|---|
| Ease of learning | 11 | N/A | 6 | 2 | 3 |
| Ease of using | 11 | N/A | 5 | 0 | 6 |
| System preference | 12 | 7 | 3 (+2)† | 0 (+2)† | 0 |

† Two liked System 2 and 3 equally.

systems different to some extent, but the level of perceived difference varied across participants.

The next two questions asked participants which of the two experimental systems they found easier to learn and use[4]. The questions were similar to the ones asked in a post-search questionnaire, but a post-test questionnaire asked participants to select either of the experimental systems or "no difference". One participant did not complete the questions due to the limited use of the hierarchy during the experiment. The result is shown in the second and third rows of Table 8.12.

As can be seen, several participants found no difference between the systems in terms of ease of learning and use of the system. This is likely to be because both systems presented terms using the same visualisation method, although an underlying technique of hierarchy generation was different. However, there appeared to be an overall preference of System 2 to System 3 among participants. This appears to correlate with the result of participants' perceptions on the hierarchies discussed in the previous section. However, the difference between two systems is not significant.

The next question asked *Which of the three systems did you like the best overall?*. The question was used to capture participants' preference of the systems. The result is shown in the bottom row of Table 8.12. As can be seen, many participants appeared to prefer System 1 to System 2 or 3. Two participants indicated that they liked System 2 and 3 equally and preferred to System 1. Therefore, several participants preferred the system with a concept hierarchy. The difference among the three systems was tested by the Friedman Test with two methods of dealing with the two participants who voted System 2 and 3. One method counted each vote as

[4]Although three systems were listed for the questions in a post-test questionnaire (See Appendix E), the control system was removed from the selection by crossing a line on system code. System A, B, and C were used as a code of the three systems in order of presentation given to a subject. Therefore, when the control system was used as the first system, System A was removed from the selection.

1, and another method counted each vote as .5. The difference is not significant in both methods.

Some of the comments given in a post-test questionnaire are as follows. Participants often commented that System 1 was simple and familiar. Two participants commented that the performance of System 1 was good enough and they did not need the hierarchy, thus, could be redundant for them. This suggests that our back-end retrieval system (Okapi system) was effective and some participants perceived that they were capable of completing a task without the support of concept hierarchies. The stronger preference to System 1 can also be due to the limited improvement made by the experimental systems in the task performance.

However, participants appeared to have cases where they found System 2 and 3 useful during a search task. Participants commented that the hierarchy gave a good idea of retrieved documents or concise view of a topic. One participant commented that s/he used the hierarchy when the search required to go deep in a ranked list to find relevant information. Other participants commented that the hierarchy was helpful for finding potential query terms. One participant commented that the hierarchy gave her/him new ideas for enriching search vocabulary. These comments seem to support our objective of supporting a user's query reformulation and searching using the concept hierarchy.

A couple of participants commented that the arrangement of terms in the hierarchy was confusing. Some participants appeared to be distracted by irrelevant terms found in the hierarchy. These comments were found in both of the experimental systems. Therefore, further development should be carried out to decrease the number of irrelevant terms populated in the hierarchy. One way to address this problem is to combine the parent-child pairs identified by multiple approaches, thus, this forms one of our future work.

### 8.5.6 Post-search learning effect

The last aspect of our investigation was regarding the topic learning effect. We devised a small task to determine if participants can formulate a better query after the 10 minutes search on the topics. We asked participants to formulate an optimised

```
After the search for a topic, we would like you to generate
a query which would retrieve the most relevant instances
based on your search experience. This query would be given
to someone who needs to search for a similar topic.
```

**Figure 8.18:** *Instruction of the query optimising task.*

**Table 8.13:** *Overall performance of the optimised queries.*

| | Initial query | Optimised query | Difference |
|---|---|---|---|
| No. of search | 36 | 36 | |
| No. of relevant docs | 4440 | 4440 | |
| No. of rel docs retrieved | 2512 | 3035 | **+523** |
| MAP | .203 | .235 | +.032 |
| P@10 | .447 | .506 | +.059 |
| P@20 | .407 | .457 | +.050 |
| R-Precision | .264 | .301 | +.037 |

query for the search topic based on their search experience. The actual instruction was shown in Figure 8.18. Participants were asked to perform this task immediately after the end of search. We recorded the optimised query and compared the retrieval effectiveness with the first query that participants submitted in the instance finding task.

We used TREC-8 Ad Hoc Track's relevance judgements data for the evaluation of the query optimising task. As described in Section 8.3.1, the topic description of the instance finding task was derived from the description of the Ad Hoc task. While the nature of the task was different, the Ad Hoc task had a deeper document pool than the Interactive Track's pool for the relevance assessments. Therefore, this allowed us to measure the performance of the traditional document-based recall and precision. The first query (called *initial* query in the results) and optimised query were submitted to the Okapi system and the top 1000 documents were retrieved for the evaluation, like the Ad Hoc task. We used a version of the `trec_eval` software developed by the InQuery team (Callan et al., 1992) which added the statistical testing function (T-Test) to the original version. In the following results, the T-Test was used to find the significant difference ($p \leq .05$) between the two types of the queries, and the significant results were highlighted in **bold**.

[*H*.10] *Participants can generate an improved query for a topic after a search session.*

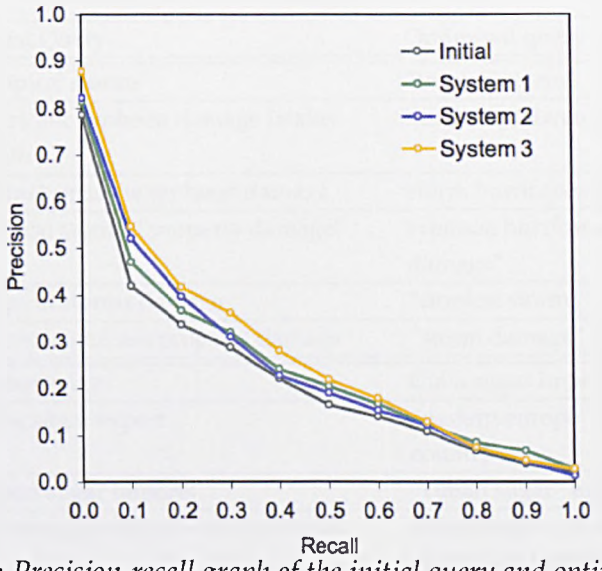



**Figure 8.19:** *Precision-recall graph of the initial query and optimised query.*

Table 8.13 shows the overall performance of the initial query and optimised query. The fourth row of the table shows that the optimised query retrieved a significantly larger number of relevant documents than the initial query. This suggests that participants were able to formulate a more effective query after the 10 minutes of search. The mean average precision (MAP), precision at Rank 10 and 20 (denoted as P@10 and P@20), and R-Precision, also show that the optimised query performed better, but their differences were not found to be significant.

The initial and optimised queries formulated by participants are shown in Table 8.14. Of 36 optimised queries, 32 had some sort of changes from the initial queries, and 4 had no change. Of the 32 changed queries, 20 had an increase in the query length, 6 had an decrease, and 6 had no difference. The number of added words in the optimised query was between one and three with the average of 1.45 words. Therefore, while the overall change made to the initial query appears to be small, participants tended to add new words more frequently than removing some words. The small changes, nevertheless, contributed to an improvement of retrieval effectiveness compared to the initial queries, increasing the number of relevant documents retrieved.

Therefore, [*H*.10] was partially supported.

[*H*.11] *Concept hierarchies can improve participants' perception on the learning of a search topic after a search session.* Figure 8.19 plots the precision-recall graph based

**Table 8.14:** *Initial and optimised queries.*

| Topic | System | Initial Query | Optimised query |
|---|---|---|---|
| 408 | 1 | "tropical storms" | "tropical storm" |
| 408 | 1 | hurricane typhoon damage fatality death | hurricane taiwan japan |
| 408 | 2 | storm hurricane typhoon damage | storm hurricanes typhoons damage |
| 408 | 2 | tropical storms "property damage" | typhoon hurricane "property damage" |
| 408 | 3 | tropical storms damage | "tropical storms" |
| 408 | 3 | hurricane names property damage | "storm damage" hurricanes typhoons |
| 414 | 1 | Cuban sugar | Cuba sugar import |
| 414 | 1 | cuba sugar export | "eastern europe" import cuba sugar country |
| 414 | 2 | cuban sugar imports | "cuban sugar" imports exports |
| 414 | 2 | cuban sugar export cane countries | cuban sugar raw exports import russian eastern europe |
| 414 | 3 | import cuban sugar | import cuban sugar cuba |
| 414 | 3 | cuba sugar imports | cuba sugar importer market |
| 428 | 1 | countries decline birth rate | countries policy decline birth rate |
| 428 | 1 | declining birth rates | declining birth rates |
| 428 | 2 | "declining birth rate" | declining birth rates |
| 428 | 2 | "birth rate" decline declining "declining birth rate" | "birth rate" fall |
| 428 | 3 | birth rate decline | birth rate decline country |
| 428 | 3 | declining birth rates countries | declining birth rates fertility rate drop countries |
| 431 | 1 | robotics | robotics achievements |
| 431 | 1 | "robotic technology" latest developments | robotics technology |
| 431 | 2 | robotic technology | robotic "new technology" |
| 431 | 2 | developments robotics | development robot technology |
| 431 | 3 | "robotic technology" | robot technology |
| 431 | 3 | robot | robot technique |
| 438 | 1 | "tourism increase" | tourism increase boom |
| 438 | 1 | increase tourism experience | tourism increase |
| 438 | 2 | tourism countries promotion | countries tourist economy |
| 438 | 2 | tourism increase | tourism increase country |
| 438 | 3 | tourism | tourist increase |
| 438 | 3 | "tourism industry" country growth | tourism growth country |
| 446 | 1 | violence against tourists | tourists hurt violence against |
| 446 | 1 | violence tourist kidnapping | violence kidnapping tourists travel warnings crime |
| 446 | 2 | tourists violence | Tourists violence |
| 446 | 2 | tourist deaths | injured tourists terrorism kidnapping |
| 446 | 3 | country violence tourism | tourism industry violence |
| 446 | 3 | "violence against tourists" | tourist killed |

**Table 8.15:** *System-breakdown of the optimised query performance.*

| System | | Initial query | Optimised query | Difference |
|---|---|---|---|---|
| 1 | MAP | .195 | .224 | +.029 |
| 2 | MAP | .208 | .228 | +.020 |
| 3 | MAP | .206 | .252 | +.046 |

on the initial and optimised query of the three systems. Since the performance of the initial query was found to be similar across the three systems, we only showed the average of the three systems for the initial query. As can be seen, the optimised query outperformed the initial query in all systems. This again confirms that participants were able to formulate a more effective query after the search. The optimised queries from System 2 and 3 tended to perform better than System 1 at the lower recall levels. The performance was switched towards the higher recall levels. While this is not a conclusive trend, there appears to be a correlation with our finding of the concept hierarchies improving the precision more than recall in the instance finding task.

An unexpected result was that System 3 tended to perform better than the other two systems, and a similar trend was found in the mean average precision shown in Table 8.15, although the difference was not significant. Our results in the instance finding task suggested that participants found System 3 less helpful for the task than System 2. We also analysed the related subjective assessment in the instance finding task. The question in the post-search questionnaire asked: *Have you learned anything new about the topic during your search?* Participants' subjective assessment on this question was similar across the systems. The mean value was 3.8 (SD: 1.4), 3.7 (SD: 1.2), and 3.9 (SD: 1.3) for System 1, 2, and 3, respectively. The difference was not significant.

Therefore, [$H$.11] was not supported.

In summary, the query optimising task was useful for confirming that participants were able to formulate an effective query after the search, since this indicated their learning effect about the topics. Similar to the instance recall and precision result, the difference between the three systems was small and insignificant. However, we also found some contradicting result compared to the instance finding task. Therefore, it was not clear if the difference of the learning effect was well captured

for the three systems by the query optimising task. Perhaps, the design of the task needs to be improved. For example, one can present the list of the terms used in the search to participants, and ask them to select (or rank) the significant terms from the list. The proposed design might be easier for participants to formulate an optimised query compared to formulating one from a scratch.

## 8.6 Discussion

The findings from our user study have several implications for the effectiveness of a concept-based approach to supporting the user's query reformulation and searching in interactive IR (IIR). The study also helped us gain further insight into the factors which were likely to affect the effectiveness and usability of the concept hierarchies in several aspect of IIR. This section discusses these implications.

One of the motivations behind the study of the automatically generated concept hierarchies was a searcher's difficulty in reformulating an effective query (Belkin, 2000). The existing ranking-based term suggestion approaches have had a mixed result of their effectiveness in the user studies (Beaulieu, 1997; Magennis and van Rijsbergen, 1997; Ruthven, 2003). It was suggested that the lack of contexts provided for the suggested terms could cause the searcher's difficulty in selecting the effective terms (Ruthven, 2003). To address this issue, our approach was to increase the awareness of the relationship among the query terms, suggested terms, and retrieved documents using the concept hierarchies. The use of the hierarchical structure also had the advantage of presenting the larger number of potentially relevant terms to the searcher, compared to the typical list presentation.

Our result shows that the concept hierarchies can help a user increase the range of search vocabulary employed to complete the task. This was a relatively consistent effect across the topics. However, the strength of the effect appeared to vary and increase in the topics where participants needed to use more terms to complete a task in the control system (System 1). The advantage of the concept hierarchies was that the user can diversify their search vocabulary without the manual effort of query reformulation. Therefore, regarding the query reformulation support, the concept hierarchies are likely to be effective when a user needs to use a range of

words to complete a task.

Another use of the concept hierarchy in our search interface was the browsing of the retrieved documents. By browsing and selecting the terms in the concept hierarchies, the experimental systems (System 2 and 3) allowed the user to focus on a particular set of retrieved documents. In other words, the experimental systems were designed to give the user a control on how the retrieved documents were grouped based on the terms in the hierarchy. In the control system, however, viewing the subsequent result pages was the only way to browse more documents retrieved by a query. Our result of the original rank position of the save documents suggests that participants found relevant information in the lower ranked documents with the concept hierarchies. This suggests that the grouping of the retrieved documents was effective to find relevant information that was not found, for example, in the first search result page. The browsing effect of the hierarchy was also found in the number of the extra result pages viewed by participants. The result shows that participants viewed a significantly fewer number of the extra result pages in the experimental systems than the control system. This suggests that participants preferred to use the concept hierarchies to explore the retrieved documents instead of viewing the subsequent result pages.

We also investigated the effect of the concept hierarchies on participants' click-through actions. Our result shows that participants made a significantly larger number of click-through in the control system than the experimental systems. The result was consistent across the topics. This appears to be the product of multiple factors. For example, participants might have spent more time on browsing the hierarchy than viewing the documents' full-texts in the experimental systems. Also, the hierarchy might have decreased the level of uncertainty in the browsing of the retrieved documents, thus, participants were able to avoid viewing perceived irrelevant documents. Based on the result of the hierarchy access over the stage of a search session and the successful click-through rate, we speculate that both factors can affect the number of click-through documents.

As can be seen, we found that the concept hierarchies can have positive effects on the different aspects of the searching process. The subsequent question was whether or not the task performance was improved by these effects. Overall, we

did not find conclusive evidence which showed that the concept hierarchies improved the performance of the instance finding task. The task performance appears to be comparable across the systems. However, our result indicates that, when an improvement is found, the hierarchy is likely to improve the precision rather than recall.

There can be several potential factors for the task performance. One was that our subjects were able to re/formulate the queries and find relevant instances without much support, since they were likely to have the advanced knowledge in the searching process and search engines compared to the general population. It was our intention to recruit such subjects to gain the critical feedback and performance measures in this study. Another was that the back-end Okapi system was effective to achieve the high precision search in response to the participants' queries. A similar observation was made in Beaulieu et al. (1999) who also used the Okapi system in a similar experiment. Another factor could be the effect of participants' mental map regarding a given topic (Wu et al., 2001). A user might have some sort of predefined mental structure about the topics. When the mental structure differs from the structure presented by the concept hierarchies, the user might find it difficult to browse the terms in the hierarchy. The disagreement between the two forms of structure can increase when the user's familiarity with the topic is high. The user's interaction with the topic familiarity is discussed in a later part of this section.

Another prominent factor appears to be the behavioural change made by the concept hierarchies. With the experimental systems, participants were likely to spend more time to browse the hierarchies and explore the retrieved documents, as opposed to a sequential browsing of the search results in the control system. While the use of the concept hierarchies had the advantages in the increase of the search vocabulary and in offering a more control on the grouping of retrieved documents, it turned out that these positive effects were not strong enough to make a significant difference in the task performance. This also suggests that the grouping of the retrieved documents should be improved. Currently, the order of the documents in the subset grouped by the hierarchy terms was the same as the original ranking. However, one can re-order the documents in the grouped result based on the selected terms, so that the result will take the user's focus into account more

significantly. This approach might improve the hierarchy's effect of "discovering" new relevant information that is hidden in the lower ranked documents.

It should be noted that the lack of significant difference in the task performance was relatively common in the Interactive Track participants (Hersh and Over, 1999). For example, the Rutgers team recruited 32 subjects in their experiment, but the difference of the task performance was not significant between their control and experimental system (Belkin et al., 1999). Therefore, there might be a case where statistical significance is found in our result of the task performance when a substantially larger number of participants are recruited in a subsequent study. However, since the overall difference of the task performance was found to be small between our control and experimental systems, it is not clear if such a subsequent study would suggest a very different result from what we have presented in this chapter.

Nevertheless, the framework provided by the Interactive Track was beneficial for us to gain further insight into the user's interaction with the concept hierarchies. For example, we found that participants' familiarity of the topics was related to the access to the hierarchy. The use of the hierarchy is likely to increase when the user is less familiar with the topic. This appears to echo the finding of Shiri and Revie (2003) who found the relationship between the familiarity and use of a thesaurus. The frequency of the hierarchy access is also likely to change within a search session. Our result of the subsumption hierarchy shows that the access to the hierarchy is likely to increase as the search progresses and the amount of new relevant information decreases, thus, the search can be perceived to be more difficult by the user. Therefore, the concept hierarchies are likely to be used more frequently in the middle to end stage of a search session than at the beginning. The analysis of the interaction path also suggests that the user is motivated to use the hierarchy when the relevant information is not found in the click-through documents.

The frequency of the hierarchy access also informed us that the keyphrase hierarchy was not used as often as the subsumption hierarchy was during the task. The subjective assessments also suggest that participants tended to find the subsumption hierarchy more useful than the keyphrase hierarchy. Participant's feedback indicates that the terms presented at the top level of the hierarchy can have a strong influence on their perception of the usefulness of the hierarchy. Therefore, the se-

lection of the top level terms was found to be an important factor to improve the effectiveness of the concept hierarchy.

Finally, we found that the level of the hierarchy accessed by participants did not differ across the stage of a search session. Our result suggests that the hierarchy does not need to be deep and the three or four levels appear to be appropriate in most cases. These findings should be taken into account for further development of the concept hierarchies.

## 8.7 Summary

This chapter presented the user study of a concept-based tool which was designed to support a user's query reformulation and searching. While the task performance was not significantly different from the control system, we found that the concept hierarchies can be beneficial for the multiple aspects of the interactive search. The study also suggested some directions to improve the effectiveness of the concept hierarchies. The next chapter discusses them in more detail.

# Chapter 9

# Conclusion and future work

## 9.1 Conclusion

The main aim of this thesis was to present a concept-based approach to supporting a user's query reformulation and searching. The approach was motivated by the findings of previous studies which suggested that the lack of context and structure in the presentation of potential candidate terms could cause difficulty in selecting appropriate terms for use query reformulation. More specifically, it was argued that the existing QE techniques did not always offer sufficient information about the relationship between the initial query terms and candidate terms, relationship between the candidate terms, and relationship between the candidate terms and the documents in which the terms were extracted. It was also pointed out that the display of term lists typically used by existing approaches were not necessarily suitable for offering a wide range of potentially useful terms in a structured way. To address this problem, we have investigated the automatic methods of generating a hierarchical structure of related concepts, and evaluated them in corpus-based analyses and user-oriented study. This section discusses the research contributions that this thesis has made.

### 9.1.1 Relationship between the document frequency and term specificity

The document frequency (DF) has been used in various IR applications. An assumption frequently made is that the DF represents a level of semantic specificity of

terms (e.g., Forsyth and Rada, 1986; Sanderson and Croft, 1999). However, limited empirical work has been carried out to support such an assumption. In Chapter 6, a large scale study was conducted to investigate the relationship between the DF and term specificity using multiple corpora and a thesaurus containing approximately 45,000 noun words and phrases.

Our experiment shows that, with 75% accuracy, DF can predict the level of specificity for a given parent-child pair in the thesaurus. However, the analysis of hypernym chains shows that this relatively high accuracy is only found at the very specific levels. Therefore, the assumption of DF representing a level of term specificity is likely to hold at the specific levels. These specific levels, nonetheless, cover the majority of vocabulary defined in the thesaurus. By looking at the shape of hypernym chains, we also observed that a high DF term can be found at the middle levels. We speculated that some terms in the peak at the middle levels might belong to the basic level categories (Rosch, 1978).

Furthermore, a larger corpus was found to be more accurate at estimating the specificity of terms than a small corpus. When the commonest sense of terms was considered, the type of corpus (i.e., web pages or news articles) did not seem to affect the effect of corpus size. However, there will be the case where it is infeasible to access to the data set as large as the index of a major search engine's collection. In such a case, the co-occurrence information was found to be effective for improving the accuracy of ordering concepts based on specificity.

Overall, this thesis presented further insight into the relationship between the DF and term specificity, which was gained by the study of multiple large corpora and the range of vocabulary extracted from a thesaurus. It also provided the empirical evidence which supported the method of exploiting the document frequency and co-occurrence information for the automatic generation of concept hierarchies.

### 9.1.2 Extraction of descriptive phrases

The previous section discussed the findings of a statistical property of terms. A distinct method investigated in this thesis was a lexical approach to finding a parent-child description based on a descriptive phrase of named entities. A lexical ap-

proach has been applied to different tasks such as populating a parent-child relationship in a thesaurus (Hearst, 1992) and finding relevant texts for a question answering system (Radev, 1998). This thesis developed and evaluated a generic tool of finding a descriptive phrase of a query noun from full-texts. The tool's design was motivated by the finding of the VLC track of TREC which suggested a simple technique could be effective for finding relevant information when a large corpus was available (Hawking and Thistlewaite, 1997).

In this thesis, the lexicon-syntactic pattern extraction approach (Hearst, 1992) was extended by an appositive phrase extraction, acronym detection, and conventional IS-A pattern extraction. The experiment showed that the performance of finding relevant descriptions improved as the size of document collection searched increased. Therefore, the result supported the finding of the VLC track. Our investigation also found that the coverage and accuracy of finding relevant descriptions varied across the extraction patterns. We proposed a weighting method which incorporated the accuracy of the patterns into the ranking of retrieved descriptions. The extended patterns were further combined with the corpus-based statistical evidences to calculate the final score of retrieved descriptions. The experiment showed that the proposed combination method of multiple evidences outperformed the effectiveness of individual evidences.

One of the observations made during the analysis of the descriptive phrases was that the phrases often provided semantically more general descriptions of the query noun. This motivated us to further investigate the effectiveness of the lexical approach as an alternative method of generating a concept hierarchy, which is discussed below.

### 9.1.3 User interaction with a concept-based search support tool

To achieve an interactive use of automatically generated concept hierarchies, we developed a search interface which was designed to present a visualised concept hierarchy to the user along with retrieved documents. A hierarchy was generated from a set of retrieved documents. To increase the awareness of the relationship among the query terms, candidate expansion terms, and retrieved documents, vi-

sual feedback was offered in the interface by highlighting terms in different colours.

We carried out a task-based user study to investigate the effectiveness of the concept hierarchies for supporting a user's query reformulation and searching. The result showed that the concept hierarchies could increase the range of search vocabulary employed to complete a search task without the manual effort of query reformulation. The result also suggested that the browsing of search results could be facilitated by the concept hierarchies. We found that the concept hierarchies could encourage a user to find relevant information in a low ranked document without viewing extra result pages. These results demonstrated that the concept hierarchies could support multiple aspects of information searching process.

The study also helped us elicit the factors that could motivate a user to use the concept hierarchies during a search task. A factor was topic familiarity. A user tended to use the hierarchy more frequently when the topic was less familiar. Another factor was the stage of search session. As the search session progressed and the probability of finding new relevant information decreased, the use of the hierarchy was likely to increase. Therefore, the frequency of the hierarchy access could increase towards the end of a search session. At a granular level of interaction, we observed that an unsuccessful click-through was likely to motivate a user to use the hierarchy. Marchionini (2006) argues the importance of making a search interface more "reactive" to a user's searching behaviour. The factors elicited in our study can be considered as the contexts which trigger the promotion of a support function to achieve such a reactive search interface.

As for the visualisation of a hierarchical structure, we found that three or four levels were sufficient for the presentation of candidate terms. A user was unlikely to browse the terms in a deep level. Also, the terms at the top level of the hierarchy appeared to influence a user's assessment of the hierarchy. A user might be discouraged to use the hierarchy when the top level terms were considered to be irrelevant to an underlying information need. Therefore, the selection of the top level terms can be important for the usability of the hierarchy in a search interface.

A limitation of the current approach was that a simple grouping of retrieved documents based on a selected term could be insufficient to achieve a significant improvement of retrieval effectiveness. Keeping the order of initial ranking in the

grouped documents could prevent a user from finding new relevant information. Therefore, there is room for improving the order of the grouped documents to improve the effectiveness of the hierarchy in interactive IR. In the following section, we discuss a potential approach to capturing a user's underlying information need from the interaction with the hierarchy, which could address the limitation of the current approach.

We believe that finding an effective way of supporting a user's information searching process is as important as the development of retrieval models and other areas in IR. This thesis demonstrated that an automatically generated concept hierarchy had an advantage of supporting multiple aspects of the searching process. Therefore, our study warrants further investigation to achieve better user support in interactive IR systems.

## 9.2 Future work

This section discusses potential direction for further investigation of the concept hierarchies. An analysis of the transaction log suggested that an area of the hierarchy that should be improved is the grouping of retrieved documents when a term was selected. Also, participants' feedback suggested there was room for improving the relationship of terms defined in a hierarchy. The following sections discuss our ideas to address the problems. Finally, we discuss an application of the concept hierarchy in a different domain.

### 9.2.1 Capturing implicit feedback via the hierarchy

Implicit feedback is a way of eliciting a searcher's underlying information need by analysing the interaction with a search interface (Kelly and Teevan, 2003). For example, display time of click-through documents has been studied as implicit feedback (Kelly and Belkin, 2004). The Ostensive Model proposes an approach to reweighting terms based on click-through documents (Campbell and van Rijsbergen, 1996). The model assumes that a searcher's information need is dynamic and developing, thus, the terms occurred in a more recently accessed document are given a

higher weight than previously accessed documents. This approach has been found to be effective to improve the retrieval effectiveness (White et al., 2004). Therefore, it appears to be worth investigating the interaction with the concept hierarchies as a means of capturing implicit feedback.

For example, the terms selected in the hierarchy can be seen as an indication of the current information need. When the grouping of retrieved documents occurs in response to the term selection, one can re-order the grouped documents based on the re-weighted selected term and existing query terms. This might improve the probability of finding relevant information in the grouped documents, compared to the current implementation which maintains the rank order of an initial retrieval. Another scenario is query expansion triggered by the term selection. Similar to the above example, one can update the weight of terms based on the interaction with the concept hierarchies. However, when term selection occurs, a search interface can submit a new query to the retrieval system using the re-weighted selected term and existing query terms. Then, the result of the new query is presented to a searcher. In this way, the search is not limited to an existing retrieved set (as our current experimental systems work), but the entire corpus is considered.

In summary, the former scenario aims to improve the effectiveness of the concept hierarchies in the browsing of retrieved documents, while the latter scenario aims to facilitate the query expansion. Both scenarios deserve further investigation since they attempt to exploit implicit feedback captured by the interaction with the concept hierarchies.

### 9.2.2 Combining multiple evidences in hierarchy formulation

This thesis investigated two approaches to generating a concept hierarchy. However, a hierarchy was generated by either of the two approaches. It seems reasonable to expect that a more robust hierarchy can be generated by using parent-child pairs derived from multiple evidence. For example, one can combine the parent-child pairs identified by the subsumption and keyphrase method to formulate a single hierarchy. Additional evidence can also be explored. For example, Bookstein et al. (2003) proposed a method of determining a symmetric or asymmetric relation-

ship between terms. Symmetricity information can be used as another evidence for hierarchy formulation.

A problem would rise when a contradicting parent-child pair is suggested by multiple sources. Therefore, a framework to resolve a contradicting case should also be considered for the combination of multiple evidences. A voting method or unified weighting scheme might be some of possible approaches to address the contradicting issue.

### 9.2.3 Concept hierarchy as a diagnostic tool

In this thesis, the concept hierarchies were mainly investigated as a search-aid tool in interactive IR. A distinct use of the automatically generated concept hierarchy can be as a diagnostic tool. In this scenario, the hierarchy can be used to offer a corpus-based evidence of the relationship between concepts. A potential area to which the diagnostic use of the hierarchy can be applied is Ontology Engineering (Corcho et al., 2003). The process of building an ontology involves much decision-makings such as the selection of key concepts in a target domain, population of other relevant concepts, and formulation of the relationship between the relevant concepts (Uschold and King, 1995). While some part of the process can be automated, the decisions are often manually finalised by a domain expert or community. Since this is an expensive task, a number of "ontology editors" have been developed to facilitate the ontology engineering (Corcho et al., 2003).

One part of the process where the concept hierarchy can be used is a bootstrapping stage. A new ontology is sometimes generated from an existing ontology or knowledge-based resource. When an appropriate resource is not available in a target domain, however, the ontology is generated from a scratch. In the latter case, an early stage of the ontology development might be facilitated by the concept hierarchies generated by a set of key concepts in the target domain. For example, a preliminary structure generated by the hierarchies might help identify frequently occurring pairs of concepts, and mine additional candidate concepts in the target domain. Another part of the process that might be facilitated by the hierarchy is the verification of the conceptual relationship defined by the ontology developer.

For a given corpus or set of documents, the hierarchy can present the relationship between the concepts based on a statistical or lexical property. Such information can be used as a corpus-based evidence to examine, verify, or update the conceptual relationships defined by the developer. The developer can also view the full-text to further clarify the context in which the concepts are used. In either case, a domain-specific corpus is likely to increase the validity of the concept hierarchy as a diagnostic tool.

To summarise, the concept hierarchy might be complementary facilitating the ontology developing process by offering corpus-based evidence, thus, further investigation in this area appears to be promising.

# References


Allan, J., Ballesteros, L., Callan, J., Croft, W., and Lu, Z. (1996). "Recent Experiments with INQUERY". In: Harman, D. K. (ed.), *Proceedings of the Fourth Text REtrieval Conference (TREC-4)*, pp. 49–72. Gaithersburg, MD: NIST.

Anick, P. (2003). "Using terminological feedback for web search refinement: a log-based study". In: Callan, J., Cormack, G., Clarke, C., Hawking, D., and Smeaton, A. (eds.), *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pp. 88–95. Toronto, Canada: ACM Press.

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

Anick, P. G. and Vaithyanathan, S. (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*, pp. 314–323. Philadelphia, PA: ACM.

Baeza-Yates, R. and Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. New York: ACM Press.

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

Beaulieu, M. (1997). "Experiments on Interfaces to Support Query Expansion". *Journal of Documentation*, **53**(1), 8–19.

Beaulieu, M. (2000). "Interaction in information searching and retrieval". *Journal of Documentation*, **56**(4), 431–439.

Beaulieu, M., Fowkes, H., Alemayehu, N., and Sanderson, M. (1999). "Interactive Okapi at Sheffield - TREC-8". In: *NIST Special Publication 500-246: The Eighth Text REtrieval Conference (TREC 8)*, pp. 689–697. Gaithersburg, MD: NIST.

Beaulieu, M. and Gatford, M. J. (1998). "Interactive Okapi at TREC-6". In: Voorhees, E. M. and Harman, D. K. (eds.), *NIST Special Publication 500-240: The Sixth Text REtrieval Conference (TREC 6)*, pp. 143–167. Gaithersburg, MD: NIST.

Belkin, N. J. (2000). "Helping people find what they don't know". *Communications of the ACM*, **43**(8), 58–61.

Belkin, N. J., Cool, C., Head, J., Jeng, J., Kelly, D., Lin, S., Lobash, L., Park, S. Y., Savage-Knepshield, P., and Sikora, C. (1999). "Relevance Feeback versus Local Context Analysis as Term Suggestion Devices: Rutgers' TREC-8 Interactive Track Experience". In: Voorheer, E. M. and Harman, D. K. (eds.), *Proceedings of the 8th Text REtrieval Conference (TREC-8)*, pp. 565–573. Gaithersburg, MD: NIST.

Belkin, N. J., Cool, C., Kelly, D., Kim, G., Kim, J.-Y., Lee, H.-J., Muresan, G., Tang, M.-C., and Yuan, X.-J. (2003). "Query Length in Interactive Information Retrieval". In: Callan, J., Cormack, G., Clarke, C., Hawking, D., and Smeaton, A. (eds.), *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 205–212. Tronto, Canada: ACM.

Bookstein, A., Kulyukin, V., Raita, T., and Nicholson, J. (2003). "Adapting Measures of Clumping Strength to Assess Term-Term Similarity". *Journal of the American Society for Information Science and Technology*, **54**(7), 611–620.

Borlund, P. (2000). "Experimental Components for the Evaluation of Interactive Information Retrieval Systems". *Journal of Documentation*, **56**(1), 71–90.

Braschler, M. and Peters, C. (2004). "Cross-Language Evaluation Forum: Objectives, Results, Achievements". *Information Retrieval*, **7**(1-2), 7–31.

Brill, E. (1992). "A simple rule-based part of speech tagger". In: Bates, M. and Stock, O. (eds.), *Proceedings of the third conference on Applied natural language processing*, pp. 152–155. Trento, Italy.

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/1921/com1921.htm` [Accessed: 6/1/07].

Callan, J. P., Croft, B. W., and Harding, S. M. (1992). "The INQUERY Retrieval System". In: *Proceedings of the Third International Conference on Database and Expert Systems Applications*, pp. 78–83. Valencia, Spain: Springer-Verlag.

Campbell, I. and van Rijsbergen, K. (1996). "The ostensive model of developing information needs". In: *Proceedings of the Second International Conference on Conceptions of Library and Information Science (COLIS-2)*, pp. 251–268. Copenhagen, Denmark.

Caraballo, S. A. and Charniak, E. (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*, pp. 63–70.

Cederberg, S. and Widdows, D. (2003). "Using LSA and Noun Coordination Information to Improve the Precision and Recall of Automatic Hyponymy Extraction". In: *Prceedings of the Seventh Conference on Natural Language Learning (CoNLL-2003)*, pp. 111–118. Edomonton, Canada: ACL.

Chen, H., Houston, A. L., Sewell, R. R., and Schatz, B. R. (1998). "Internet browsing and searching: User evaluations of category map and concept space techniques". *Journal of the American Society for Information Science*, **49**(7), 582–608.

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.

Corcho, O., Fernández-López, M., and Gómez-Pérez, A. (2003). "Methodologies, tools and languages for building ontologies. Where is their meeting point?" *Data and Knowledge Engineering*, **46**(1), 41–64.

Crestani, F., Lalmas, M., van Rijsbergen, C. J., and Campbell, I. (1998). ""Is this document relevant? ... probably": a survey of probabilistic models in information retrieval". *ACM Computing Surveys*, **30**(4), 528 – 552.

Croft, B. W. and Lafferty, J. (eds.) (2003). *Language Modeling for Information Retrieval*, volume 13 of *Kluwer International Series on Information Retrieval*. Kluwer Academic Publishers.

Cutting, D. R., Pedersen, J. O., Karger, D., and Tukey, J. W. (1992). "Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections". In: Belkin, N. J., Ingwersen, P., and Pejtersen, A. M. (eds.), *Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 318–329. Copenhagen, Denmark: ACM.

Deboeck, G. and Kohonen, T. (eds.) (1998). *Visual Explorations in Finance with Self-Organizing Maps*. Berlin: Springer.

Doszkocs, T. E. (1983). "CITE NLM: Natural-Language Searching in an Online Catalog". *Information Technology and Libraries*, **2**, 364–380.

Efthimiadis, E. N. (1993). "A user-centred evaluation of ranking algorithms for interactive query expansion". In: Korfhage, R., Rasmussen, E. M., and Willet, P.

(eds.), *Proceedings of the 16th Annual ACM SIGIR conference on Research and Development in Information Retrieve*, pp. 146–159. Pittsburgh, PA: ACM.

Efthimiadis, E. N. (1996). "Query Expansion". *Annual Review of Information Systems and Technology*, **31**, 121–187.

Efthimiadis, E. N. (2000). "Interactive Query Expansion: A User-Based Evaluation in a Relevance Feedback Environment". *Journal of the American Society for Information Science*, **51**(11), 989–1003.

Enser, P., Kompatsiaris, Y., and O'Connor, N. E. (eds.) (2004). *Image and Video Retrieval: Third International Conference (CIVR 2004)*, volume 3115. Dublin, Ireland: Springer-Verlag. Lecture Notes in Computer Science 3115.

Forsyth, R. and Rada, R. (1986). "Adding an Edge". In: *Machine Learning: applications in expert systems and information retrieval*, pp. 198–212. Ellis Horwood.

Fowkes, H. and Beaulieu, M. (2000). "Interactive searching behaviour: Okapi experiment for TREC 8". In: Robertson, S. and Ayse, G. (eds.), *Proceedings of the BCS-IRSG 22nd Annual Colloquium on Information Retrieval Research*, pp. 47–56. Cambridge, UK: BSC-IRSG.

Fox, S., Karnawat, K., Mydland, M., Dumais, S., and White, T. (2005). "Evaluating implicit measures to improve web search". *ACM Transactions on Information Systems*, **23**(2), 147–168. 1059982.

Frakes, W. B. (1992). "Stemming algorithms". In: Frakes, W. B. and Baeza-Yates, R. (eds.), *Information retrieval: data structures and algorithms*, pp. 131–160. Upper Saddle River, NJ: Prentice-Hall.

Fuhr, N., Lalmas, M., and Malik, S. (eds.) (2004). *Proceedings of the Second Annual Workshop of the Initiative for the Evaluation of XML Retrieval (INEX)*. Dagstuhl, Germany: ERCIM.

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: Webber, B. L. (ed.), *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL-EACL 2001)*, pp. 196–203. Toulouse, France.

Gibson, D., Kleinberg, J. M., and Raghavan, P. (1998). "Inferring Web Communities from Link Topology". In: *Proceedings of the 9th ACM Conference on Hypertext and Hypermedia*, pp. 225–234. Pittsburgh, Pennsylvania: ACM.

Glover, E., Pennock, D. M., Lawrence, S., and Krovetz, R. (2002). "Inferring hierarchical descriptions". In: *Proceedings of the Eleventh International Conference on Information and Knowledge Management (CIKM'02)*, pp. 507–514. McLean, VA: ACM.

Greenberg, J. (2001a). "Automatic Query Expansion via Lexical-Semantic Relationships". *Journal of the American Society for Information Science and Technology*, **52**(5), 402–415.

Greenberg, J. (2001b). "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.

Grefenstette, G. (1992). "Use of syntactic context to produce term association lists for retrieval". In: Belkin, N. J., Ingwersen, P., and Pejtersen, A. M. (eds.), *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 89–97. Copenhagen, Denmark: ACM.

Grefenstette, G. (1997). "SQLET: Short Query Linguistic Expansion Techniques, Palliating One-Word Queries by Providing Intermediate Structure to Text". In: *Proceedings of Recherche d'Informations Assistee par Ordinateur - Computer Assisted Information Retrieval (RIAO '97)*, pp. 500–509.

Grefenstette, G. and Tapanainen, P. (1994). "What is a word, What is a sentence? Problems of Tokenisation". In: Kiefer, F. K., Kiss, G., and Pajzs, J. (eds.), *Proceedings of the 3rd Conference on Computational Lexicography and Text Research (COMPLEX'94)*, pp. 79–87. Budapest, Hungary: Hungarian Academy of Sciences.

Hancock-Beaulieu, M., Fieldhouse, M., and Do, T. (1995). "An Evaluation of Interactive Query Expansion in an Online Library Catalogue with a Graphical User Interface". *Journal of Documentation*, **51**(3), 225–243.

Harman, D. (1988). "Towards Interactive Query Expansion". In: *Proceedings of the 11th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 321–331. Grenoble, France: ACM.

Harman, D. (1992). "Overview of the First Text REtrieval Conference (TREC-1)". In: Harman, D. K. (ed.), *NIST Special Publication 500-207: The First Text REtrieval Conference (TREC-1)*, pp. 1–20. Gaithersburg, MD: NIST.

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

Hearst, M. A. (1992). "Automatic Acquisition of Hyponyms from Large Text Corpora". In: *Proceedings of the Fourteenth International Conference on Computational Linguistics*, pp. 539–545. Nantes, France.

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. Cambridge, MA: MIT Press.

Hearst, M. A. (1999). "User Interfaces and Visualization". In: Baeza-Yates, R. and Ribeiro-Neto, B. (eds.), *Modern Information Retrieval*, pp. 257–323. New York: ACM Press.

Hearst, M. A. and Pederson, J. O. (1996). "Re-examining the Cluster Hypothesis: Scatter/Gather on Retrieval Results". In: Frei, H.-P., Harman, D., Schable, P., and Wilkinson, R. (eds.), *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 76–84. Zurich, Switzerland: ACM.

Hersh, W. and Over, P. (1999). "TREC-8 Interactive Track Report". In: Voorheer, E. M. and Harman, D. K. (eds.), *NIST Special Publication 500-246: The Eighth Text REtrieval Conference (TREC-8)*, pp. 57–64. Gaithersburg, ML: NIST.

Hersh, W., Turpin, A., Price, S., Kraemer, D., Olson, D., Chan, B., and Sacherek, L. (2001). "Challenging conventional assumptions of automated information retrieval with real users: Boolean searching and batch retrieval evaluations". *Information Processing & Management*, **37**(3), 383–402.

Ingwersen, P. (1992). *Information Retrieval Interaction*. London: Taylor Graham Publishing.

Jansen, B. J., Spink, A., Bateman, J., and Saracevic, T. (1998). "Real Life Information Retrieval: A Study of User Queries on the Web". *ACM SIGIR Forum: A Publication of the Special Interest Group on Information Retrieval*, **32**(1), 5–17.

Joho, H., Coverson, C., Sanderson, M., and Beaulieu, M. (2002). "Hierarchical Presentation of Expansion Terms". In: Lamont, G. B., Haddad, H., Papadopoulos, G., and Panda, B. (eds.), *Proceedings of the 17th ACM Symposium on Applied Computing (SAC'02)*, pp. 645–649. Madrid, Spain: 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.

Jones, S., Gatford, M., Robertson, S., Hancock-Beaulieu, M., and Secker, J. (1995). "Interactive Thesaurus Navigation: Intelligent Rules OK?" *Journal of the American Society for Information Sicence*, **46**(1), 52–59.

Kekäläinen, J. and Järvelin, K. (1998). "The impact of query structure and query expansion on retrieval performance". In: Croft, B. W., Moffat, A., van Rijsbergen, C. J., Wilkinson, R., and Zobel, J. (eds.), *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 130–135. Melbourne, Australia: ACM.

Kelly, D. and Belkin, N. J. (2004). "Display time as implicit feedback: understanding task effects". In: *Proceedings of the 27th annual international conference on Research and development in information retrieval*, pp. 377–384. Sheffield, United Kingdom: ACM Press. 1009057.

Kelly, D. and Teevan, J. (2003). "Implicit feedback for inferring user preference: a bibliography". *SIGIR Forum*, **37**(2), 18–28.

Kobayashi, M. and Takeda, K. (2000). "Information retrieval on the web". *ACM Computing Surveys*, **32**(2), 144–173.

Koenemann, J. and Belkin, N. J. (1996). "A Case for Interaction: A Study of Interactive Information Retrieval Behavior and Effectiveness". In: *Conference Proceedings on Human Factors in Computing Systems (CHI '96)*, pp. 205–212. Vancouver, Canada: ACM.

Kohonen, T. (1995). *Self-Organizing Maps*. Berlin: Springer.

Kristensen, J. (1993). "Expanding End-Users' Query Statements for Free Text Searching with a Search-Aid Thesaurus". *Information Processing & Management*, **29**(6), 733–744.

Krovetz, R. (1993). "Viewing morphology as an inference process". In: Korfhage, R. (ed.), *Proceedings of the 16th International Annual ACM SIGIR Conference on Research and Development of Information Retrieval*, pp. 191–202. Pittsburgh: ACM.

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

Magennis, M. and van Rijsbergen, C. J. (1997). "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*, pp. 324–332. Philadelphia, PA: ACM.

Marchionini, G. (2006). "Leveraging Context for Adaptive Multimedia Retrieval: A Matter of Control". In: Detyniecki, M., Jose, J. M., Nürnberger, A., and van Rijsbergen, C. J. (eds.), *Adaptive Multimedia Retrieval: User, Context, and Feedback, Third International Workshop (AMR 2005)*, Lecture Notes in Computer Science, 3877, pp. 35–43. Glasgow, UK: Springer.

Miller, G. A. (1990). "Nouns in WordNet: A Lexical Inheritance System". *International Journal of Lexicography*, **3**(4), 245–264.

Miller, G. A. (1995). "WordNet: A Lexical Database for English". *Communications of the ACM*, **38**(11), 39–41.

Nanas, N., Uren, V., and De Roeck, A. (2003). "Building and applying a concept hierarchy representation of a user profile". In: Callan, J., Cormack, G., Clarke, C., Hawking, D., and Smeaton, A. (eds.), *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 198–204. Tronto, Canada: ACM.

Niwa, Y., Nishioka, S., Iwayama, M., and Takano, A. (1997). "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)*, pp. 95–100. Phuket, Thailand.

Over, P. (1997). "TREC-6 Interactive Report". In: Voorheer, E. M. and Harman, D. (eds.), *NIST Special Publication 500-240: The Sixth Text REtrieval Conference (TREC 6)*, pp. 73–81. Gaithersburg, MD: NIST.

Over, P. (1998). "TREC-7 Interactive Track Report". In: Voorheer, E. M. and Harman, D. (eds.), *NIST Special Publication 500-242: The Seventh Text REtrieval Conference (TREC 7)*, pp. 33–39. Gaithersburg, MD: NIST.

Palmer, D. D. and Hearst, M. A. (1997). "Adaptive Multilingual Sentence Boundary Disambiguation". *Computational Linguistics*, **23**(2), 241–267.

Paynter, G. W. and Witten, I. H. (2001). "A combined phrase and thesaurus browser for large document collections". In: *Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries, Lecture Notes In Computer Science, Vol. 2163*, pp. 25–36.

Peat, H. J. and Willet, P. (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.

Pollitt, S. (1997). "The key role of classification and indexing in view-based searching". In: *Proceedings of the 63rd IFLA General Conference*. Copenhagen, Denmark. Available from `http://www.ifla.org/IV/ifla63/63cp.htm` [Accessed: 6/1/07].

Ponte, J. M. and Croft, B. W. (1998). "A Language Modeling Approach to Information Retrieval". In: Croft, B. W., Moffat, A., van Rijsbergen, C. J., Wilkinson, R., and Zobel, J. (eds.), *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 275–281. Melbourne, Australia: ACM.

Porter, M. and Galpin, V. (1988). "Relevance feedback in a public access catalogue for a research libraries: Muscat at the Scott Polar Research Institute". *Program*, **22**(1), 1–20.

Porter, M. F. (1980). "An algorithm for suffix stripping". *Program*, **14**, 130–137.

Qiu, Y. and Frei, H. P. (1993). "Concept Based Query Expansion". In: Korfhage, R., Rasmussen, E. M., and Willett, P. (eds.), *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 160–169. Pittsburgh, PA: ACM.

Radev, D. R. (1998). "Learning correlations between linguistic indicators and semantic constraints: Reuse of context-dependent descriptions of entities". In: *Proceedings of the Joint 17th International Conference on Computational Linguistics 36th Annual Meeting of the Association for Computational Linguistics (COLING-ACL'98)*, pp. 1072–1078. Montreal, Canada: ACL.

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

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

Robertson, S. E. and Sparck Jones, K. (1976). "Relevance weighting of search terms". *Journal of the Americal Society for Information Science*, **27**(3), 129–146.

Robertson, S. E., Walker, S., and Beaulieu, M. (1997). "Laboratory Experiments with Okapi: Participation in the TREC Programme". *Journal of Documentation*, **53**(1), 20–34.

Robertson, S. E., Walker, S., and Beaulieu, M. (1998). "Okapi at TREC-7: automatic ad hoc, filtering VLC and interactive track". In: Voorheer, E. M. and Harman, D. K. (eds.), *NIST Special Publication 500-242: The Seventh Text REtrieval Conference (TREC-7)*, pp. 253–264. Gaithersburg, MD: NIST.

Robertson, S. E., Walker, S., Beaulieu, M. M., Gatford, M., and Payne, A. (1995b). "Okapi at TREC-4". In: Harman, D. K. (ed.), *NIST Special Publication 500-236: The Fourth Text REtrieval Conference (TREC-4)*, pp. 73–97. Gaithersburg, MD: NIST.

Rosch, E. (1978). "Principles of Categorization". In: Rosch, E. and Lloyd, B. B. (eds.), *Cognition and Categorization*, pp. 27–48. Lawrence Erlbaum Associates.

Ruthven, I. (2003). "Re-examining the Potential Effectiveness of Interactive Query Expansion". In: Callan, J., Cormack, G., Clarke, C., Hawking, D., and Smeaton, A. (eds.), *Proceedings of the 26th Annual International ACM SIGIR Conference on*

*Research and Development in Information Retrieval*, pp. 213–220. Tronto, Canada: ACM.

Ryu, P.-M. and Choi, K.-S. (2006). "Determining the specificity of terms using inside-outside information: a necessary condition of term hierarchy mining". *Information Processing Letters*, **100**(2), 76–82.

Salton, G. (1968). *Automatic Information Organization and Retrieval.* New York: McGraw-Hill.

Salton, G. (1971). *The SMART Retrieval System.* Englewood Cliffs, NJ: Prentice-Hall.

Salton, G. (1980). "Automatic term class construction using relevance. A summary of work in automatic pseudoclassification". *Information Processing & Management*, **16**, 1–15.

Salton, G. and Buckley, C. (1988). "Term-Weighting Approaches in Automatic Text Retrieval". *Information Processing & Management*, **24**(5), 513–523.

Salton, G. and Buckley, C. (1990). "Improving Retrieval Performance by Relevance Feedback". *Journal of the American Society for Information Science*, **41**(4), 288–297.

Salton, G. and McGill, J. (1983). *Introduction to Modern Information Retrieval.* McGraw-Hill.

Sanderson, M. and Croft, B. (1999). "Deriving Concept Hierarchies from Text". In: Hearst, M., Gey, G., and Tong, R. (eds.), *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 206–213. Berkeley, CA: ACM.

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

Schatz, B. R., Johnson, E. H., Cochrane, P. A., and Chen, H. (1996). "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*, pp. 126–133. Bethesda, MD: ACM.

Schütze, H. and Pedersen, J. O. (1997). "A Cooccurrence-based Thesaurus and Two Applications to Information Retrieval". *Information Processing & Management*, **33**(3), 307–318.

Sebastiani, F. (2001). "Interactive Query Expansion with Automatically Generated Category-Specific Thesauri". In: Chin, A. G. (ed.), *Text Databases and Document Management: Theory and Practice*, pp. 103–117. Hershey, US: Idea Group Publishing.

Sebrechts, M. M., Vasilakis, J., Miller, M. S., Cugini, J. V., and Laskowski, S. J. (1999). "Visualization of Search Results: A Comparative Evaluation of Text, 2D, and 3D Interfaces". In: Hearst, M., Gey, G., and Tong, R. (eds.), *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 3–10. Berkeley, CA: ACM.

Shiri, A. A. and Revie, C. (2003). "The effects of topic complexity and familiarity on cognitive and physical moves in a thesaurus-enhanced search environment". *Journal of Information Science*, **29**(6), 517–526.

Siegel, S. and Castellan, J. N. (1988). *Nonparametric Statistics for the Behavioral Sciences.*. McGraw-Hill, 2nd edition.

Silverstein, C., Henzinger, M., Marais, H., and Moricz, M. (1998). *Analysis of a Very Large AltaVista Query Log.* Technical Note 1998-14, Systems Research Center (SRC). Available from `http://gatekeeper.dec.com/pub/DEC/SRC/technical-notes/abstracts/src-tn-1998-014.html` [Accessed: 6/1/07].

Sparck Jones, K. (1971). *Automatic Keyword Classification for Information Retrieval.* London: Butterworths.

Sparck Jones, K. (1972). "A Statistical Interpretation of Term Specificity and its Application in Retrieval". *Journal of Documentation*, **28**(1), 11–21.

Sparck Jones, K., Walker, S., and Robertson, S. E. (2000a). "A probabilistic model of information retrieval: development and comparative experiments Part 1". *Information Processing & Management*, **36**(6), 779–808.

Sparck Jones, K., Walker, S., and Robertson, S. E. (2000b). "A probabilistic model of information retrieval: development and comparative experiments Part 2". *Information Processing & Management*, **36**(6), 809–840.

Sparck Jones, K. and Willett, P. (eds.) (1997). *Readings in Information Retrieval.* San Francisco, CA: Morgan Kaufmann.

Spink, A., Wolfram, D., Jansen, M. B. J., and Saracevic, T. (2001). "Searching the Web: The Public and Their Queries". *Journal of the American Society for Information Science and Technology*, **52**(3), 226–234.

Uschold, M. and King, M. (1995). "Towards a Methodology for Building Ontologies". In: *Proceedings of the Workshop on Basic Ontological Issues in Knowledge Sharing at the International Joint Conference on Artificial Intelligence (IJCAI-95).* Montreal, Canada. Available from `http://www.aiai.ed.ac.uk/publications/tr95.html` [Accessed: 6/1/07].

Van Rijsbergen, C. J. (1979). *Information Retrieval.* London: Butterworths, 2nd edition.

Voorhees, E. M. (1994). "Query Expansion Using Lexical-Semantic Relations". In: Croft, B. W. and Rijsbergen, J. C. v. (eds.), *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 61–69. Berlin, Germany: ACM.

Voorhees, E. M. and Buckland, L. P. (eds.) (2004). *NIST Special Publication 500-261: The Thirteenth Text REtrieval Conference Proceedings (TREC 2004)*. Gaithersburg, MD: NIST.

Voorhees, E. M. and Harman, D. (1997). "Overview of the Sixth Text REtrieval Conference (TERC-6)". In: Voorhees, E. and Harman, D. (eds.), *NIST Special Publication 500-240: The Sixth Text REtrieval Conference (TREC 6)*, pp. 1–24. Gaithersburg, MD: NIST.

Voorhees, E. M. and Harman, D. K. (1998). "Overview of the Seventh Text REtrieval Conference (TREC-7)". In: Voorhees, E. M. and Harman, D. K. (eds.), *NIST Special Publication 500-242: The Seventh Text REtrieval Conference (TREC 7)*, pp. 1–24. Gaithersburg, MD: NIST.

Wacholder, N., Evans, D. K., and Klavans, J. L. (2001). "Automatic identification and organization of index terms for interactive browsing". In: *Proceedings of the first ACM/IEEE-CS joint conference on Digital libraries*, pp. 126–134. Roanoke, VA: ACM.

Wacholder, N. and Nevill-Manning, C. G. (2001). "The Technology of Phrase Browsing Applications: Workshop held in conjunction with the first ACM-IEEE joint conference on digital libraries". *SIGIR Forum*, **35**(1), 16–20.

Wade, S. J. and Willett, P. (1988). "INSTRUCT: a teaching package for experimental methods in information retrieval. Part III. Browsing, clustering and query expansion". *Program*, **22**(1), 44–61.

Weinberg, B. H. and Cunningham, J. A. (1985). "The Relationship Between Term Specificity in MeSH and Online Postings in MEDLINE". *Bulletin of the Medican Library Association*, **73**(4), 365–372.

White, R. W., Jose, J. M., van Rijsbergen, C. J., and Ruthven, I. (2004). "A Simulated Study of Implicit Feedback Models". In: McDonald, S. and Tait, J. (eds.), *Advances in Information Retrieval, 26th European Conference on IR*, pp. 311–326. Sunderland, UK.

White, R. W., Ruthven, I., and Jose, J. M. (2005). "A study of factors affecting the utility of implicit relevance feedback". In: *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 35–42. Salvador, Brazil: ACM.

Wiesman, F., van den Herik, H. J., and Hasman, A. (2004). "Information retrieval by metabrowsing". *Journal of the American Society for Information Science and Technology*, **55**(7), 565–578.

Wise, J. A., Thomas, J. J., Pennock, K., Lantrip, D., Pottier, M., and Schur, A. (1995). "Visualizing the non-visual: Spatial analysis and interaction with information from text documents". In: *Proceedings of the IEEE Information Visualization Symposium (InfoVis '95)*, pp. 51–58. Atlanta, GA: IEEE.

Wittgenstein, L. (1953). *Philosophical Investigations*. New York: Macmillan.

Woods, W. A. (1997). *Conceptual indexing: a better way to organize knowledge*. Technical Report TR-97-61, Sun Microsoft Laboratories.

Wu, M., Fuller, M., and Wilkinson, R. (2001). "Using clustering and classification approaches in interactive retrieval". *Information Processing & Management*, **37**(3), 459–484.

Xu, J. and Croft, B. W. (1996). "Query Expansion Using Local and Global Document Analysis". In: Frei, H.-P., Harman, D., Schauble, P., and Wilkinson, R. (eds.), *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 4–11. Zurich, Switzerland: ACM.

Xu, J. and Croft, B. W. (2000). "Improving the Effectiveness of Information Retrieval with Local Context Analysis". *ACM Transactions on Information Systems*, **18**(1), 79–112.

# Appendices

# Appendix A

# Instruction for participants

This instruction was given to participants of the user study presented in Chapter 8.

# Scenario

Imagine that you have just returned from a visit to your doctor during which it was discovered that you are suffering from high blood pressure. The doctor suggests that you take a new experimental drug, but you wonder what alternative treatments are currently available. You decide to investigate the literature on your own to satisfy your need for information about what different alternatives are available to you for high blood pressure treatment. You really need only one document for each of the different treatments for high blood pressure.

You find and save a single document that lists four treatment drugs. Then you find and save another two documents that each discusses a separate alternative treatment: one that discusses the use of calcium and one that talks about regular exercise. You've run out of time and stop your search. In all, you have identified six different instances of alternative treatments in three documents.

---

In this experiment, you will face a similar task. You will be presented with several descriptions of needed information on a number of topics. In each case there can be multiple examples or instances of the type of information that's needed.

We would like you to identify as many different instances as you can of the needed information for each topic that will be presented to you - as many as you can in the 10 minutes you will be given to search. Please save one document for EACH DIFFERENT instance of the needed information that you identify. If you save one document that contains several instances, try not to save additional documents that contain ONLY those instances. However, you will not be penalized if you save documents unnecessarily.

As you identify an instance of the needed information, please keep track of which instances you have found: write down a word or short phrase to identify the instance, or--if the system provides a facility to keep track of instances--use it.

Carefully read each topic to understand the type of information needed. This will vary from topic to topic. On one topic you may be looking for instances of a certain kind of event. On another you may be searching for examples of certain sorts of people, places, or things.

Do you have any questions about

- what we mean by instances of needed information
- the way in which you are to save non-redundant documents for each instance?"

---

After the search for a topic, we would like you to generate a query which would retrieve the most relevant instances based on your search experience. This query would be given to someone who needs to search for a similar topic.

March 2003

# Appendix B

# Entry questionnaire

This questionnaire was designed to capture participants' background information at the beginning of the experiment presented in Chapter 8.

## ENTRY QUESTIONNAIRE

**Background Information**

1. What high school/college/university degrees/diplomas do you have (or expect to have)?

| Degree | Major | Date |
|---|---|---|
| Degree | Major | Date |
| Degree | Major | Date |
| Degree | Major | Date |
| Degree | Major | Date |

2. What is your occupation?

_____

3. What is your gender?

____ Female ____ Male

4. What is your age?

___ 18 – 27 years
___ 28 – 37 years
___ 38 – 47 years
___ 48 – 57 years
___ 58 – 67 years
___ 68+

**Computer Experience**

Please circle the number that most closely describes your computer experience.

| How much experience have you had... | None | | | Some | | | A great deal |
|---|---|---|---|---|---|---|---|
| 1. using computers? | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 2. using World Wide Web browsers? | 1 | 2 | 3 | 4 | 5 | 6 | 7 |

| How often do you use a computer for... | Never | | | Monthly | | | Daily |
|---|---|---|---|---|---|---|---|
| 1. work tasks? | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 2. academic tasks? | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 3. personal tasks? | 1 | 2 | 3 | 4 | 5 | 6 | 7 |

Please indicate your level of expertise with computers:

| Novice | | | | | | Expert |
|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 |

**Searching Experience**

Please indicate the number that most closely describes your searching experience.

| How much experience have you had... | None | | | Some | | | A great deal |
|---|---|---|---|---|---|---|---|
| 1. searching with WWW search engines? | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 2. searching with online library catalogs? | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 3. searching with indexing/abstracting service (INSPEC, LISA, etc.)? | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 4. searching with other systems, please specify the system: | | | | | | | |
| a._____ | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| b._____ | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| c._____ | 1 | 2 | 3 | 4 | 5 | 6 | 7 |

| 5. How much knowledge do you have of how search engines work. | None | | | Some | | | A great deal |
|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 6. When I search the WWW, I can usually find what I am looking for. | Rarely | | | Some-times | | | Often |
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |

| How often do you conduct searching for information about... | Never | | | Monthly | | | Daily |
|---|---|---|---|---|---|---|---|
| 1. assignment/work related project? | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 2. shopping? | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 3. traveling? | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 4. medical/health? | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 5. government policy? | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 8. entertainment? | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 9. other information, please specify: | | | | | | | |
| a._____ | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| b._____ | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| c._____ | 1 | 2 | 3 | 4 | 5 | 6 | 7 |

Please indicate your level of expertise with searching:

| Novice | | | | | | Expert |
|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 |

Overall, for how many years have you been doing online searching? _____ years

Please list your favorite search engine(s): _____.
_____.
_____.

Have you ever participated in a TREC experiment? **YES / NO**

# Appendix C

# Post-search questionnaire (Control System)

This questionnaire was designed to capture participants' subjective assessments of search, and used after the search session of the control system.

## POST-SEARCH QUESTIONNAIRE

Please indicate the number that most closely describes your opinions.

| | Not at all | | | Some-what | | | Extremely |
|---|---|---|---|---|---|---|---|
| 1. Was it easy to get started on this search? | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 2. Was it easy to do the search on this topic? | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 3. Are you satisfied with your search results? | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 4. Did you have enough time to do an effective search? | 1 | 2 | 3 | 4 | 5 | 6 | 7 |

| | None | | | Some | | | A great deal |
|---|---|---|---|---|---|---|---|
| 5. Did your previous knowledge help you with your search? | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 6. Have you learned anything new about the topic during your search? | 1 | 2 | 3 | 4 | 5 | 6 | 7 |

Please write any other comments about the task for this particular topic.

# Appendix D

# Post-search questionnaire (Experimental System)

This questionnaire was designed to capture participants' subjective assessments of search, and used after the search session of the experimental systems.

## POST-SEARCH QUESTIONNAIRE

Please indicate the number that most closely describes your opinions.

| | Not at all | | | Some-what | | | Extremely |
|---|---|---|---|---|---|---|---|
| 1. Was it easy to get started on this search? | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 2. Was it easy to do the search on this topic? | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 3. Are you satisfied with your search results? | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 4. Did you have enough time to do an effective search? | 1 | 2 | 3 | 4 | 5 | 6 | 7 |

| | None | | | Some | | | A great deal |
|---|---|---|---|---|---|---|---|
| 5. Did your previous knowledge help you with your search? | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 6. Have you learned anything new about the topic during your search? | 1 | 2 | 3 | 4 | 5 | 6 | 7 |

| | Not at all | | | Some-times | | | Always |
|---|---|---|---|---|---|---|---|
| 7. Was it easy to browse the hierarchical menu? | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 8. Was the menu too deep/long to manage? | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 9. Was the menu confusing or misleading? | 1 | 2 | 3 | 4 | 5 | 6 | 7 |

| | Not at all | | | Some-times | | | Always |
|---|---|---|---|---|---|---|---|
| 10. Were the terms in the menu helpful for predicting the contents of the linked documents? | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 11. Were the terms in the menu helpful for judging the relevance of documents? | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 12. Was the menu helpful in focusing on important terms? | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 13. Were the terms in the menu helpful for understanding the retrieved documents? | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 14. By browsing the menu, did you feel that you had a better idea of the contents of a set of retrieved documents? | 1 | 2 | 3 | 4 | 5 | 6 | 7 |

Was there any terms you did not expect to find in the menu? And examples?

Please write any other comments about the task for this particular topic.

# Appendix E

# Post-test questionnaire

This questionnaire was designed to capture participants' subjective assessments of system preference and other feedback, and used at the end of three search sessions.

## POST-TEST QUESTIONNAIRE

How different did you find the systems from one another?

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| not at all | | somewhat | | extremely |

Which of the two systems did you find easier to learn?

❑ System A ❑ System B ❑ System C ❑ No difference

Which of the two systems did you find easier to use?

❑ System A ❑ System B ❑ System C ❑ No difference

Which of the three systems did you like the best overall?

❑ System A ❑ System B ❑ System C ❑ No difference

What did you like about each of the systems?

System A

System B

System C

Please list any other comments that you have about your overall search experience.

That's it! Thank you very much for your time.

# Appendix F

# Sample transaction log

This was a snippet of the transaction log recorded during the experiment presented in Chapter 8.

```
#
# CiQuest Usability Test Log File
#
#    Session ID: test001_menu_subsump
#      Topic ID: 428i
#  Generated at: Fri Mar 14 11:53:17 2003
#
<log>
 <session_id>test001_menu_subsump</session_id>
 <topic_id>428i</topic_id>
 <date>Fri Mar 14 11:53:17 2003</date>
 <action type="new">
  <time>11:53:17</time>
 </action>
 <action type="submit">
  <time>11:54:50</time>
  <info>
   <query>"declining birth rate"</query>
  </info>
 </action>
 <action type="spell">
  <time>11:54:50</time>
  <info>
   <suggested></suggested>
  </info>
 </action>
 <action type="search">
  <time>11:54:51</time>
  <info>
   <hitnum>11</hitnum>
   <query>
    <oneline>"declining birth rate"</oneline>
    <no>1</no>
     <term id="1">
      <nostem>"declining birth rate"</nostem>
      <stem>declin birth rate</stem>
      <df>11</df>
      <weight>9.813</weight>
     </term>
   </query>
  </info>
 </action>
 <action type="show">
  <time>11:54:53</time>
  <info>
   <start>1</start>
   <end>10</end>
   <doc>
    <record rank="1">
     <intDocID>176595</intDocID>
     <extDocID>FT943-5357</extDocID>
    </record>
    <record rank="2">
     <intDocID>153575</intDocID>
     <extDocID>FT941-16778</extDocID>
    </record>
    <record rank="3">
     <intDocID>180861</intDocID>
     <extDocID>FT943-9226</extDocID>
```

```
    </record>
    <record rank="4">
     <intDocID>175248</intDocID>
     <extDocID>FT943-4010</extDocID>
    </record>
    <record rank="5">
     <intDocID>174296</intDocID>
     <extDocID>FT943-193</extDocID>
    </record>
    <record rank="6">
     <intDocID>183006</intDocID>
     <extDocID>FT943-11018</extDocID>
    </record>
    <record rank="7">
     <intDocID>109178</intDocID>
     <extDocID>FT933-7393</extDocID>
    </record>
    <record rank="8">
     <intDocID>123582</intDocID>
     <extDocID>FT934-5273</extDocID>
    </record>
    <record rank="9">
     <intDocID>122597</intDocID>
     <extDocID>FT934-4288</extDocID>
    </record>
    <record rank="10">
     <intDocID>135</intDocID>
     <extDocID>FT911-135</extDocID>
    </record>
   </doc>
  </info>
</action>
<action type="expand">
 <time>11:55:14</time>
 <info>
  <term>birth</term>
  <df>10</df>
  <location>1_1_1_1</location>
  <doc>
   <record rank="1">
    <intDocID>176595</intDocID>
    <extDocID>FT943-5357</extDocID>
   </record>
   <record rank="2">
    <intDocID>153575</intDocID>
    <extDocID>FT941-16778</extDocID>
   </record>
   <record rank="3">
    <intDocID>180861</intDocID>
    <extDocID>FT943-9226</extDocID>
   </record>
   <record rank="4">
    <intDocID>175248</intDocID>
    <extDocID>FT943-4010</extDocID>
   </record>
   <record rank="5">
    <intDocID>174296</intDocID>
    <extDocID>FT943-193</extDocID>
   </record>
```

```
    <record rank="6">
     <intDocID>183006</intDocID>
     <extDocID>FT943-11018</extDocID>
    </record>
    <record rank="7">
     <intDocID>109178</intDocID>
     <extDocID>FT933-7393</extDocID>
    </record>
    <record rank="8">
     <intDocID>123582</intDocID>
     <extDocID>FT934-5273</extDocID>
    </record>
    <record rank="9">
     <intDocID>122597</intDocID>
     <extDocID>FT934-4288</extDocID>
    </record>
    <record rank="10">
     <intDocID>135</intDocID>
     <extDocID>FT911-135</extDocID>
    </record>
   </doc>
  </info>
</action>
<action type="view">
 <time>11:55:18</time>
 <info>
  <intDocID>176595</intDocID>
  <extDocID>FT943-5357</extDocID>
  <rank>0</rank>
 </info>
</action>
<action type="save">
 <time>11:55:24</time>
 <info>
  <intDocID>176595</intDocID>
  <extDocID>FT943-5357</extDocID>
  <rank>0</rank>
 </info>
</action>
<action type="view">
 <time>11:55:34</time>
 <info>
  <intDocID>153575</intDocID>
  <extDocID>FT941-16778</extDocID>
  <rank>0</rank>
 </info>
</action>
<action type="save">
 <time>11:55:42</time>
 <info>
  <intDocID>153575</intDocID>
  <extDocID>FT941-16778</extDocID>
  <rank>0</rank>
 </info>
</action>
<action type="view">
 <time>11:55:47</time>
 <info>
  <intDocID>180861</intDocID>
```