ABSTRACT

One of the main difficulties that users of information retrieval (IR) systems face is how to choose appropriate search terms to express their information need in the form of a query. In order to assist searchers, interactive query expansion (IQE) is a feature provided by IR systems, which involves suggesting to users terms that may be worth adding to their query; searchers can then interactively select some of these terms to expand their initial query. Research has shown that the uptake of IQE is limited, because searchers are unsure as to why expansion terms have been suggested and how they relate to their initial query. Searchers are also unsure as to what effect the selection of some of the expansion terms will have on the search results.

In this study, we address the second issue by providing searchers with additional information in order to support their expansion decisions. More specifically, we generate summaries of the results that would be retrieved if the searchers selected any combinations of expansion terms to add to their initial query. The summaries are query-biased, and aim to display the context in which the selected query terms would appear in the retrieved results if the searcher was to select the respective expansion terms. We evaluate our approach through user experiments where we compared our system with a baseline system with no expansion and a system where IQE was provided with no summaries. We investigate whether the query-biased summaries of retrieval results are effective at supporting searchers complete their information seeking tasks. We also focus on difficult and unfamiliar tasks and study whether there is an uptake in the use of suggested expansion terms when searchers reformulate their queries.

The pattern of data collected from our user study suggests that providing users with additional information to predict the effect of suggested expansion terms on their results lead to an increase in the use of expansion terms for query reformulations. We report a 16% increase in the use of suggested expansion terms for query reformulations across all tasks. For unfamiliar tasks, the increase in the use of suggested expansion terms was more noticeable at 35%. Searchers also found that suggested terms were most intuitive when summaries were provided and they understood how to use the terms. There is therefore strong evidence that summaries provide additional support to searchers in making expansion decisions especially for hard and unfamiliar tasks.

Table of Contents

Chapter 1

INTRODUCTION

The problem of retrieving information has been apparent for a long time. It can be dated to the days of Plato (Van Rijsbergen, 1995).

"And how will you enquire, Socrates, into that which you did not know?...if you find what you want, how will you ever know that this is the thing which you did not know?...a man cannot enquire either about that which he knows, or about that which he does not know: for if he knows, he has no need to enquire; and if not, he cannot; for he does not know the very subject about which he is to enquire" – Plato

To this day, some of the issues in this paradoxical quote persist in the area of information retrieval (IR). Searchers find it difficult to identify the correct words to formulate their information needs and if their queries are not phrased appropriately, there is likely to be a negative impact on the results they receive from the search system. Therefore, to ensure that users are able to get the best set of results that meet their information needs, it is of utmost importance to provide assistance to users of IR systems.

Neither the user nor the system is able to solve the information problem independently. Therefore, they should complement each other (Brajnik et al., 2002). The system should provide certain features to assist users, especially those users who have little or no knowledge about the area of their search. Searchers are often not able to exploit features being provided to them by the system. Sometimes when searchers assess retrieved documents, they are not able to fully exploit the information they collect. For example, when searchers judge a document as being relevant, they could extract some terms from that document to better reformulate their query in order to retrieve more similar documents (Marchionini, 1989; Sullivan et al., 1990; Vollaro & Hawkins, 1986). One of the main difficulties that users of IR systems face is how to choose appropriate search terms to express their information need in the form of a query.

One of the ways to assist searchers is through Query Expansion, in which an IR system tries to improve user queries by adding related terms to the original query. The new formulation of the query, with the added expansion terms, aims at improving the effectiveness of the search.

Query Expansion exists in two types, namely Automatic Query Expansion (AQE) and Interactive Query Expansion (IQE).

Automatic query expansion is where the retrieval system automatically adds words to the query and the user has no control over which terms have been added during the search (Efthimiadis, 2000). Ruthven (Ruthven, 2003) argues that AQE is often favourable as the system has access to more statistical information on how useful expansion terms could be. Thus, the system can select the terms that would be best for the results.

Interactive query expansion, on the other hand, is the case where the user indicates which expansion terms should be added to their query. Searchers are usually presented with a list of terms generated by the system and they can choose terms which are relevant to their query. The user-selected terms are then added to the initial query. This type of expansion is often preferred by users as it offers them more control. Koenemann and Belkin (Koenemann and Belkin, 1996) provide evidence that searchers may prefer the interactive method regardless of whether it improves performance.

1.1 Motivation

Despite the searchers' preference for interactive query expansion, a number of studies, for example, (Magennis and Van Rijsbergen, 1997) and (Ruthven, 2003), have shown that the uptake of IQE is limited, despite it having good potential. The situation is such mainly because of the following two reasons.

Firstly, searchers are unsure as to why expansion terms have been suggested to them and how they relate to their initial query. Despite numerous assumptions (Magennis and Van Rijsbergen, 1997), research (Blocks et al., 2002) has shown that searchers can fail to make the most of their understanding of semantics between their search area and suggested expansion terms, especially when they do not have background knowledge in their area of search. This shows that despite humans having the intuitive ability to exploit semantic relationships, they need the system to explicitly provide them with evidence of such relationships.

Secondly, searchers are also unsure as to what effect the selection of some of the expansion terms will have on their search results (White and Marchionini, 2006). This is so because search systems currently do not provide users with any kind of assistance in this area. Searchers are encouraged to use expansion words to reformulate their queries but they are still uncertain as to how this reformulation of their initial query will help them to retrieve better results for their search task. We are unaware of any attempt to address this issue and there has been no research activity, to the best of our knowledge, which studies how the potential of interactive query expansion could be exploited if searchers were provided with a context that allowed them to perceive how suggested expansion words could improve their results.

Therefore, this project will address the second issue. The study will focus on providing additional information to searchers so that they are able to see the impact that their choice of expansion words has on the search results. There is a clear and apparent need to provide assistance to users of interactive query expansion systems so that they can make good expansion decisions. We believe that there is greater need of support for users in an *'unknown environment'*[1], to help them exploit the potential of interactive query expansion in their search tasks.

[1] In this report, *'unknown environment'* will be used to refer to circumstances where the searcher does not have prior knowledge of the search area, that is, they do not have background information about their topic to help them make informed expansion decisions.

1.2 Aims and Objectives

The aim of this project is to support users of interactive query expansion in making good expansion decisions to improve search effectiveness, especially in areas of little or no, previous knowledge. Users will be provided with additional information in the form of query-biased summaries, to help them determine how their choice of expansion words will affect their search results.

This project has the following objectives:

- To perform background research on areas of relevance feedback, with particular attention to query expansion and interface design for interactive information retrieval.
- Build a retrieval system to be able to retrieve documents in response to a query and generate a list of expansion terms for a given query.
- Design and implement an interface to allow searchers to select the expansion terms and view retrieved documents.
- Create a feature that allows searchers to view the impact of their choice of expansion terms on the results.
- Investigate through user experiments, whether the query biased summary help users to make good expansion decisions.
- Analyse data collected from user experiments to assess performance of designed solution in supporting different types of users in different situations.

1.3 Report Structure

The first chapter, *Introduction*, outlines the motivation behind the project. It includes the project's aims and objectives.

The second chapter, *Background Research*, contains the literature review. In this section, I describe the different approaches taken by previous work and how my work fits in.

The third chapter, *Design and Implementation*, describes the different software and algorithms which have been used to develop the search system with an interactive query expansion feature.

The fourth chapter, *Comparison of the effectiveness of the expansion algorithms*, includes the methodology and findings from the comparison of the Rocchio weighting scheme and the wpq weighting algorithm.

The fifth chapter, *User Study: Design,* explains how the laboratory-based study was conducted. It includes details of the systems, tasks and data collection methods for the experiments.

The sixth chapter, *User Study: Results and Discussion*, reports on the findings from our user experiments. It includes a discussion on the different questions being address.

The final chapter, *Conclusion*, contains the concluding comments from the study and outlines the scope for future work.

Chapter 2

BACKGROUND RESEARCH

Over the years a number of studies have concentrated on query expansion. Research has focussed on a number of issues, ranging from term selection and weighting to comparisons between the different types of query expansion. In this section, different topics related to query expansion, for example, relevance feedback and interface support, are discussed and critically analysed.

2.1 Relevance Feedback

Relevance feedback is often seen as a way to involve users in the information retrieval process with the aim of improving the final result set. Users give feedback on the relevance of documents in an initial set of results. Relevance feedback exploits the idea that it may be difficult to formulate a good query when searchers do not know the content of the collection well. However, as it is easy to judge the relevance of particular documents, it makes sense to engage in the kind of iterative refinement that relevance feedback is about (Manning et al., 2008).

(White et al., 2005) describe two types of relevance feedback, namely, implicit relevance feedback (IRF) and explicit relevance feedback (ERF). Implicit relevance feedback is the process by which a search system gathers evidence on the users' interest. Information about users' search behaviour is extracted from the way they interact with the system. This is currently being exploited by search engines like Google. Users' interests have been modelled by numerous methods, both on a short term and a long term basis. Long-term modelling methods range from analysing the user's personal document collection (Chirita et al. 2007 and Dengel, 2006) to observing the user's actions like visited websites (Teevan et al., 2005). Short term modelling methods investigate users' immediate information needs and take into account actions such as scrolling, viewing and querying behaviour (Kelly and Teevan, 2003).

Explicit relevance feedback allows users to mark documents in the result set as relevant and tend to be used in situations where the system is not able to exploit the statistics of user interaction with the system. The study by White et al. (White et al., 2005) showed that IRF was more effective and useful for search tasks with high complexity and that users' preference for IRF tends to increase with task complexity. Also, IRF was found to be more efficient and easy by inexperienced users, but they still felt more in control with ERF.

Although experienced users tend to feel comfortable with explicit relevance feedback, the technique is often not popular among searchers as shown in (Ruthven et al., 2001). In this study, the authors used document summaries to allow searchers to provide relevance feedback information more quickly. Despite users liking the simplicity of the summaries, relevance feedback was not popular because users could not understand the mechanism of how it works and also because they felt that the interface to allow them to make relevance

feedback decisions increased the effort required on their behalf, as well as the cognitive load imposed on them.

Another form of relevance feedback is pseudo relevance feedback (PRF) which is also known as blind relevance feedback (Manning et al., 2008). In this form of relevance feedback, the top n, for example, 5 or 10 documents from the result set are assumed to be relevant by the system and are used to improve future results. In a number of studies (White and Marchionini, 2006; Ruthven, 2003), PRF has been used to generate terms for query expansion.

2.2 Query Expansion

While relevance feedback is an interactive tool allowing users to indicate which documents are relevant in a retrieved set, query expansion is often described as the technique where users of a search system are provided with a shortlist of suggested terms in order to enhance the effectiveness of searches (Harman, 1988). There are two types of query expansion:

- ***Automatic Query Expansion (AQE)***
  Additional terms are added to the initial query by the system with no user involvement.

- ***Interactive Query Expansion (IQE)***
  Users are presented with a list of suggested expansion terms and they are left to make their own choices as to which words they want to add to their query.

2.2.1 Generating terms for query expansion

There are different ways of deriving the list of expansion words namely, relevance feedback, nearest neighbour and term variants. In the relevance feedback method, the list of terms is generated from the documents that the searcher designates as relevant. Nearest neighbour is a technique which involves the use of a manually-built thesaurus (Voorhees, 94), an option which is most often not easily available (Harman, 1988). Expansion words are derived using the synonyms of words in the query. Term variants, on the other hand, are different variations of one word in the original query, and these are used to generate the expansion list. For example, the term variants for the word *"brightness"* could be *"bright"* and *"brightly"*.

Harman (Harman, 1988) studied which of the three different techniques improved query expansion the most; while relevance feedback was found to be costly in terms of time, as relevant terms have to be assembled, merged and ranked, the research showed that term variants and nearest neighbours tend to transform a query into a set of more productive words as these techniques are independent of the number of relevant documents retrieved at the first pass. Yet, the experiments demonstrated that applying term variants and nearest neighbour and filtering only terms that appear in relevant documents produced the most effective results.

2.2.1.1 Local and Global Analysis

Global analysis and local analysis are two approaches for generating expansion terms for query expansion. The former can be defined as a set of techniques that analyse the corpus of documents to establish relationships between words (Xu and Croft, 1996). These methods involve expanding or reformulating queries with no dependency on the original query and results returned from it (Manning et al., 2008).

On the other hand, local analysis works by adjusting the query relative to the documents that have originally been retrieved when the initial query was first run (Manning et al., 2008). Local methods, therefore, are highly dependent on the proportion of relevant documents and undoubtedly the performance of local techniques is impacted by the number of relevant documents and whether there are any relevant documents at all. Thus, these methods are known to be highly unpredictable and often not effective.

Xu and Croft (Xu and Croft, 1996) discuss a new technique identified as Local Context Analysis, whereby, ideas from global analysis, for example, context and phrase structure, are borrowed and applied to the context of local techniques. This has been very effective as shown in their study.

2.2.2 Automatic and interactive query expansion

There have been a number of studies conducted to compare AQE and IQE. For example, Koenemann and Belkin (Koenemann and Belkin, 1996) showed how IQE can outperform AQE for certain tasks and Beaulieu (Beaulieu, 1997) demonstrated that AQE produced more effective results in an operational environment. However, as mentioned in (Ruthven, 2003), the discrepancy in findings results from the difference in methodology, user experiments, search tasks and interface design. Ruthven (Ruthven, 2003) demonstrated how a comparison between query expansion and no query expansion resulted in 50% of improved queries when AQE was used. However, when he studied AQE and IQE, he found that it was difficult for users to make good IQE decisions in order to outperform the AQE decisions of a system. Therefore, it seems like the potential benefits of IQE are not being exploited.

It is difficult for searchers to realise the effectiveness of IQE without training and experience in terms of selection and search strategies. In reality, IQE is based on the hypothesis that it is reasonable to assume that searchers will be able to distinguish between good and bad terms, if they are given a list of terms (Magennis and Van Rijsbergen, 1997). This underlying assumption was however proved wrong in studies where experiments involving real searchers to use IQE failed to show any significant improvement on AQE (Araya, 1990 and Hancock-Beaulieu et al., 1995). Yet, there is evidence that users prefer to use IQE even if it does not improve performance because they feel more in control (Koenemann and Belkin, 1996).

Therefore, it appears the IQE could improve search effectiveness, but users have difficulties in exploiting its potential. Interface design seems to be one of the reasons why searchers cannot achieve the real benefits of IQE (Ruthven, 2003). Simple term listing interfaces are not adequate to provide support in making good expansion decisions. Interfaces should aim

at supporting the identification of relationship between relevant material and suggested expansion terms (Ruthven, 2003).

2.3 Interface Support for Interactive Information Retrieval

Interactive query expansion requires increased effort on behalf of users because it involves added user interface functions (Magennis and Van Rijsbergen, 1997). Interactive information retrieval imposes on the users a decision making task and also forces them to think and have a strategy. These underlying arguments can be justified as interactive systems imply that some input is needed from the user and in these situations, the user is aware their interaction with the system will influence the output they get from the system. In the context of interactive query expansion, the decision making and strategy task can be related to deciding which terms to select and which to reject in expanding their initial query. While making these decisions, they are aware that the choices they make are going to have a direct impact on the results of their search and this forces them to think in a strategic way, as ultimately, the aim of any searcher is to be able to find the most relevant set of results which meet their information needs (Magennis and Van Rijsbergen, 1997).

For this reason, interfaces supporting interactive IR increase the cognitive load on users. Thus, it is important that systems supporting interactive functions such as interactive query expansion (IQE) provide the necessary support as outlined by White and Ruthven (White and Ruthven, 2006). Ruthven (Ruthven, 2003) discusses how it is difficult to achieve the full potential of IQE because term presentation interfaces do not support users in making good expansion decisions.

This issue will be addressed in this project, to make sure users can exploit the IQE feature provided in the best of ways. Searchers will be able to see in real time on a dedicated area of the interface, how selected expansion words will affect their search results. White and Marchionini in (White and Marchionini, 2006) use a real time approach known as Real-Time Query Expansion (RTQE) and suggest that it increased the uptake of query expansion. By showing a clear relationship between relevant information and suggested terms, we aim to address what Beaulieu and Jones (Beaulieu and Jones, 1998) referred to as '*functional visibility*' relating to interfaces with a query expansion feature. They suggest that there should be a feeling of awareness among searchers about the features provided by the system. However, searchers should know about the effect of the options available to them.

2.4 Research Objectives

In examining the potential effectiveness of interactive query expansion, Ruthven (Ruthven, 2003) recognises the need to provide more sophisticated support so that searchers are able to assess the potential quality of expansion words. Previous work (Magennis and Van Rijsbergen, 1997) has shown that users have difficulty in exploiting the effectiveness of interactive query expansion.

The principal question asked in this project, is the reason behind searchers not making use of a feature that could help improve their search results. It should have been an intuitive response on the side of users to exploit the use of functionalities that allow them to retrieve

information that matches their information needs best. As already discussed in section 1.1, there are several reasons behind this problem. Thus, the main aim of this project is to examine whether users would be more likely to exploit the potential effectiveness of interactive query expansion if they had more information about how different selections of expansion words would affect their search results.

As already mentioned in (Ruthven, 2003), users need additional support to be able to make good expansion decisions. In this project, the additional support provided to users, will be in the form of *retrieved-document-set (RDS)* summaries. The objective here is to implement a predictive feature to give users an idea of the kind of documents that will be retrieved, if they choose to add one or more suggested expansion terms to their initial query.

Retrieved-document-set (RDS) summaries in this report will refer to the query-biased summary which has been generated using the set of documents that is retrieved for one particular search. Tombros and Sanderson refer to query-biased summaries as being user directed, whereby summaries are customised to reflect the information need expressed by the user in the new query (Tombros and Sanderson, 1998). While the above mentioned study uses summarisation for one document at a time, in this project, summarisation will be carried on a document-set level because the aim is to allow searchers to have a glimpse of what the retrieved document set will contain, should they choose to add certain expansion terms which has been suggested by the system.

Chapter 3

DESIGN AND IMPLEMENTATION

A search system was implemented with an IQE feature where searchers can see the effect that their choice of expansion words have on their results. The system consists of the following components: a search engine, an expansion term generator, a summariser and document indexes. This chapter outlines the design and implementation details of this IQE system.

3.1 System Overview

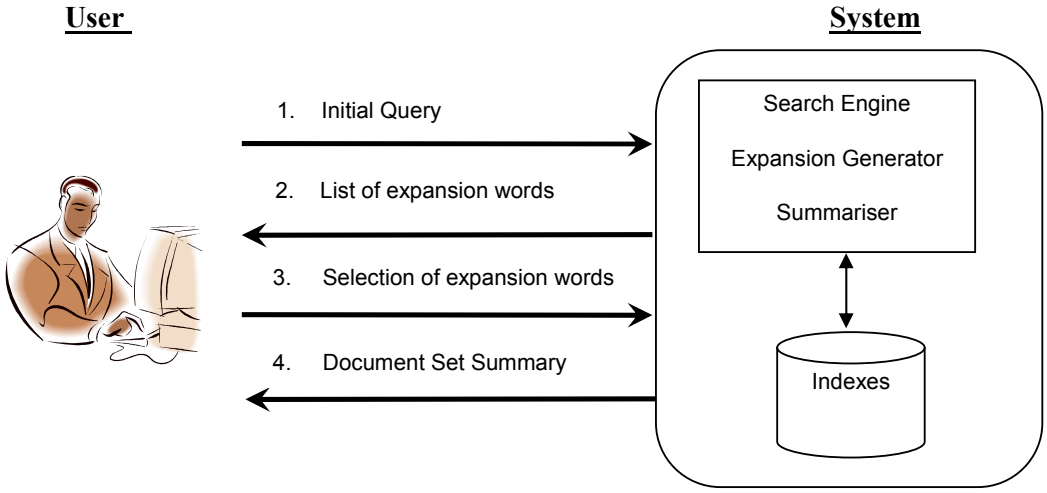


Figure 1: System overview of implemented search system

Figure 1 shows how users interact with the system through a search interface and submit their initial query. The search engine retrieves a set of documents for the initial query and produces a list of expansion words, that is, additional terms that could be added to the query to potentially increase the effectiveness of the search. The user decides which terms to select from the list. For each selection that the user makes, the system generates a summary of the document set that would be retrieved, should the user decide to add these expansion words to their initial query. The user then has the choice to resubmit the new query, which includes the initial query, selected expansion words and any other terms that the searcher might choose to add manually.

3.2 Document Collection

The documents contained in the index were articles from the Wall Street Journal (WSJ) for the year 1990 to 1992 and the Associated Press (AP) for the year 1988, taken from the TREC-5 collection (Text Retrieval Conferences) (Voorhees and Harman, 1995). Table 1 provides additional information about the collections being used in this system.

	Number of documents	Median number of terms/ document	Mean number of terms/document
Wall Street Journal, 1990-1992	74, 520	301	508.4
Associated Press, 1988	79, 919	438	468.7

Table 1: Document Collection Statistics from TREC-5

TREC collections are typically used for IR experiments and the documents are in an SGML format with tags enclosing different parts of one document. The documents are arranged in a structure that allows easy parsing. Figure 2 shows an example of a document from the WSJ collection and Table 1 contains the parts of the documents that are used in this system.

```
<DOC>
<DOCNO>
WSJ900516-0177
</DOCNO>
<DOCID>
900516-0177.
</DOCID>
<HL>
   Apache Discovers Oil Off Texas
</HL>
<DATE>
05/16/90
</DATE>
<SO>
WALL STREET JOURNAL (J), PAGE C6
</SO>
<CO>
   APA UNP
</CO>
<LP>
   DENVER, Colo. -- Apache Corp. said it discovered gas in
the Gulf of Mexico, offshore Texas.
   The well flowed at a rate of 14.1 million cubic feet of
gas a day through a 23/64-inch choke at a depth of 7,300
feet.
</LP>
<TEXT>
   Apache is the operator with a 40% interest. Its partners
are Union Pacific Resources Co. of Fort Worth, Texas, a unit
of Union Pacific Corp. of New York with a 35% interest, and
Santa Fe International of Dallas, with a 25% interest.
   ---
   Corrections &amp;amp; Amplifications


   UNION PACIFIC Corp. is based in Bethlehem, Pa. The
location was incorrect in yesterday's edition.
   (WSJ May 17, 1990)
</TEXT>
</DOC>
```

Figure 2: Example article from the document collection

Tag	Description
<DOCNO>	Unique identifier for the document
<HL>	Head Line of the article
<DATE>	Date the article was published
<SO>	The source of the article
<LP>	The leading paragraph of the article
<TEXT>	The body of the article

Table 2: Description of tags used in system

3.3 The Information Retrieval System

3.3.1 Text pre-processing

Lucene[2] is a scalable Information Retrieval library that allows users to add indexing and searching capabilities to their applications. It is not a search engine in itself, but provides a framework for applications requiring a search function (Gospodnetic, 2005).

In this project, Lucene was used to create an index from the document collection. However, before the articles could be indexed, they had to be converted into a suitable structure from their SGML format. The articles were therefore parsed into different sections, namely, *headline*, *date*, *source*, *leading paragraph* and *text* according to their respective tags. A Lucene index is made up of *Document* objects having a number of attributes called fields. In this system, one *Document* object was created for each article from the collection and had the following fields: headline, date, source, leading paragraph and text.

The *Documents* were analysed and indexed by the Lucene *Analyzer*. Analysis in Lucene is the process of converting field text into its most fundamental indexed representation, terms. Terms are used to determine which documents match a query during searches. The *Analyzer* tokenises text by performing a number of operations such as extracting words, discarding punctuations, removing accents etc. Lucene provides different implementations for the *Analyzer*, such as the *StopAnalyzer* where the user has the option of supplying the list of stop words, they would like to be removed from the text being indexed. Figure 3 shows how one sentence is analysed by the Lucene *Analyzer*.

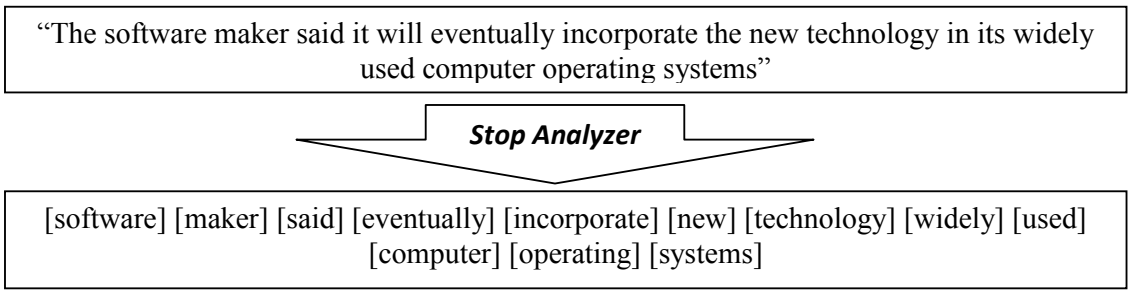


Figure 3: Analysing a sentence with the *Stop Analyzer*

[2] http://www.lucene.apache.org

This system also consists of a sentence index, which includes each sentence of every document in the document index. To build the sentence index, the articles from the document collection were split into their respective sentences. The Sentence Detector tool in the OpenNLP[3] Project was used to perform this task. OpenNLP is a set of java-based Natural Language Processing (NLP) tools which perform sentence detection, tokenisation, chunking and parsing, named-entity detection, and co reference.

3.3.2 Document Scoring

A retrieval status value (RSV) for s document (d) in the index is computed based on the query (q) using the following formula:

$$\sum_{t\ in\ d} tf(t\ in\ d).idf(t).boost(t.field\ in\ d).lengthNorm(t.field\ in\ d).\text{coord(q, d)}.\text{queryNorm(q)}$$

Factors	Description
tf (t in d)	Term frequency factor for the term *(t)* in the document *(d)*, that is, the number of times term *t* appears in the document *d*.
idf (t)	Inverse document frequency of the term *t*. The *idf* measures the importance of term *t* in the document collection. It is calculated by dividing the number of documents in the collection by the number of documents containing *t*.
boost (t in field in d)	Some fields in the document, for example, the title the article, can be more important than others. Thus, they are given a higher weight. The default boost value for a field is 1.
lengthNorm (t field in d)	Normalization by the total number of terms in the field. Normalisation value is $\frac{1}{\sqrt{Number of\ terms\ in\ field}}$
coord(q, d)	Coordination factor, based on the number of query terms the document contains. The coordination factor gives an AND-like boost to documents that contain more of the search terms than other documents.
queryNorm(q)	Given the sum of the squared weights of each of the query terms, the normalisation value for the query is $\frac{1}{\sqrt{sum of\ weights\ of\ query\ terms}}$

[3] http://opennlp.sourceforge.net

3.4 Expansion Terms Generator

This section describes the component of the system responsible for generating a list of 10 expansion terms to display to searcher after they submit their initial query to the system. The expansion terms are displayed in alphabetical order on the interface. Rudimentary stemming is applied to remove plurals. For example, if the term '*bank'* and *'banks'* are both present in the list of potential expansion terms, we remove the term *'banks'*.

3.4.1 Generating expansion terms

As mentioned in 2.2.1, there are several approaches to generating potential expansion words. Relevance feedback was one of the methods identified to derive terms for query expansion. However, it is not always popular with users as they are usually hesitant to provide feedback explicitly, or they do not want to lengthen the interaction with the search system (Manning et al, 2008).With users of the Excite search engine, Spink et al. show that only about 4% of user query sessions used the relevance feedback options (Spink et al., 2000).

Therefore, in this study, pseudo relevance feedback is used to generate potential expansion words. Also known as blind relevance feedback, this approach provides an alternative to the manual part of relevance feedback, where users have to explicitly indicate which documents are relevant. This method of local analysis allows the users to have improved retrieval performance without an extended interaction (Manning et al., 2008). An initial set of documents is retrieved for a query and the top 10 documents are assumed to be relevant and relevance feedback is applied as before. Pseudo relevance feedback was used with the Cornell SMART system (Buckley et al., 1995) at TREC 4 and produced very competitive results in various TREC competitions.

3.4.2 Weighting expansion terms

It is necessary to present the expansion words to the user in a reasonable order, that is, terms which are more likely to be useful should be higher in the list (Robertson, 1990). In this study, terms will be ranked in descending order of their weight, implying that the highest ranked term is likely to be the most useful. The methods of weighting and ranking terms investigated were the Rocchio weight (Rocchio, 1971) and the *wpq* weight (Robertson, 1990).

The Rocchio weight is derived from the Rocchio relevance feedback algorithm, whereby queries are reformulated based on a set of previously identified relevant documents. Query terms and terms in those relevant documents are reweighted according to the following formula:

$$Q' = \alpha Q + \beta \left(\frac{1}{R'} \sum_{i \, \varepsilon \, D_{R'}} DOC_i \right)$$

Q'	New improved query vector, starting from Q
Q	Initial Query
α , β	Constants, set to 1.0 and 0.5 respectively, where α is the weight for terms in the initial query and β is the weight for terms in relevant documents.
R'	Number of relevant documents
$D_{R'}$	Set of R' documents identified as relevant
DOC_i	The vector for the i[th] document

LucQE[4] is a Lucene module which performs query expansion based on the above formula.

In figure 4, we demonstrate how the formula works in practice. We have an initial query Q with the terms {account, bank} given weights {5, 3} respectively. We also have a relevant document R, which contains terms {account, bank, money, savings} and their weights are {3, 2, 1, 1}

	Terms				
	account	bank	money	savings	source
Initial Query, Q	5	3	0	0	0
Relevant Document, R,	3	2	1	1	0

If we construct a new query Q' using the Rocchio formula, we get the following:

$$Q' = 1(5,3,0,0,0) + 0.5(3,2,1,1,0)$$

$$Q' = (5,3,0,0,0) + (1.5,1,0.5,0.5,0)$$

$$Q' = (6.5,4,0.5,0.5,0)$$

Figure 4: Weighting terms with the Rocchio algorithm

Therefore the new query now contains all the four terms present in the relevant documents, namely {account, bank, money, savings}. The weight for the existing query terms {bank, account} has also increased as their presence in the initial query and the relevant document means they are important terms. Two new terms {money, savings} have been added to Q' as they were present in the relevant documents. If we were generating expansion terms from this example, {money, savings} would have been part of the list of suggested terms.

[4] http://lucene-qe.sourceforge.net/

Our implementation of a second type of weighting and ranking algorithm is the *wpq* algorithm (Robertson, 1990). The following formula is used:

$$wpq_t = w_t(p_t - q_t)$$

where w_t is the weighting function, p_t is the probability that term t occurs in a relevant document and q_t is the probability that term t occurs in a non relevant document.

This algorithm implies that irrespective of the weighting function used, *wpq* is an effective way of ranking terms because it factors the fact that high frequency terms improve recall[5] and very infrequent terms improve precision[6], by adding$(p_t - q_t)$. The complete *wpq* formula is therefore shown below.

$$wpq_t = \log \frac{r_t/(R - r_t)}{(n_t - r_t)/(N - n_t - R + r_t)} \bullet \left(\frac{r_t}{R} - \frac{n_t - r_t}{N - R}\right)$$

r_t	Number of seen relevant documents containing term t
n_t	Number of documents containing term t
R	Number of seen relevant documents
N	Number of documents in the collection

In our implementation of the *wpq* algorithm, a pseudo relevance feedback approach was adopted, that is, the number of seen relevant documents was assumed to be 10. Candidate terms were extracted from the relevant documents and stop words were removed. To avoid falling to the 'curse of dimensionality' (Van Rijsbergen, 1979), we tried different approaches to limit the number of terms that were being weighted and ranked as we were aware that from the list of extracted words, some would be more useful than others. Moreover, weighting a large number of potential expansion terms might not be efficient as we only use ten of the highest ranked for query expansion.

Our first approach was to use the leading paragraph of each document to extract potential terms instead of using the full text itself. This attempt was derived from the work of White et al. (White et al., 2007) which used text snippets to generate expansion terms. The similarity between query expansion and pseudo-relevance feedback was found to be higher when titles or text snippets were used rather than full text. Thus, our approach included using the leading paragraph instead of the full text of the article. Our decision was based on the fact that the leading paragraph would contain words that exist in similar context as the initial query, so our algorithm would not slow the generation of expansion terms and the probability of deviating from the query would be lower (White et al., 2007). However, as

[5] Recall measures the proportion of relevant documents retrieved [Number of relevant documents retrieved divided by the total number of relevant documents in the collection] (Salton and McGill, 1983).
[6] Precision measures the proportion of retrieved items that are relevant [Number of relevant documents divided by total number of documents retrieved] (Salton and McGill, 1983).

will be reported in section 4.2, the performance of the *wpq* algorithm deteriorated with this approach.

Therefore, our next approach was to use the *tf-idf* score of the extracted terms to limit the number of terms being weighted. As the *tf-idf* score gives a measure of the term's importance in the document corpus, we ranked all extracted terms by the tf-idf score. The 50 terms with the highest score were then passed to the wpq weighting algorithm. Fifty terms were chosen as we felt that it would still provide a good set of terms, even after existing query terms were removed from the list. In case, the list of extracted terms was less than 50, all of the extracted terms were weighted by our algorithm. This approach was more efficient in terms of time as there was no need to weight terms which will not be useful for expansion.

The effect of these three approaches on the performance of the *wpq* algorithm is discussed in Chapter 4, which reports a comparison between the Rocchio and the *wpq* algorithms. In figure 5, we show how terms are weighted with the *wpq* algorithm.

Assuming the initial query *'bank account'* was submitted to the system and the following are some terms found in the top 10 documents: {savings, money, currency, online, banking}. In our example, we will explain how the weight of the term *'savings'* and *'online'* will be calculated using the *wpq* algorithm.

	Term : '*savings*'
$r_{savings}$	5
$n_{savings}$	45
R	10
N	300

	Term : '*online*'
$r_{savings}$	2
$n_{savings}$	10
R	10
N	300

$$wpq_{savings} = \log\frac{5/(10-5)}{(45-5)/(300-45-10+5)} \bullet \left(\frac{5}{10} - \frac{45-5}{300-10}\right) = 0.29$$

$$wpq_{online} = \log\frac{2/(10-2)}{(10-2)/(300-10-10+2)} \bullet \left(\frac{2}{10} - \frac{10-2}{300-10}\right) = 0.16$$

Figure 5: Weighting terms with *wpq* algorithm

As the term *'savings'* appear in more of the relevant documents that the term *'online'*, it has a higher weight. Therefore, it would rank higher in the list of potential expansion terms.

3.5 Summariser

The main aim of this study is to provide searchers with a context to see what effect their choice of expansion terms has on the search results. The summary builder of the search system helps to achieve this objective. The system computes summaries on the fly,

depending on the searcher's actions. If the searcher chooses to add one term to their query, reads the summary and decides to add another term, the summary is built and updated in real time. The same applies if the searcher chooses to remove an expansion term from the list of selected terms.

The system adds the chosen term to the searcher's initial query and retrieves a set of documents *(NewQueryDocs)* matching this new query. We used the top 25 documents to build the summary as good results have been reported with IQE when 25 documents were used to generate expansion terms (Ruthven, 2003). The summary is made up of 5 sentences. This value seems to agree with the suggestions in (Edmundson, 1964) and (Brandow et al., 1995). We tried two approaches to generating the summary, a global approach and a local approach, which are both graphically represented in Figure 6.

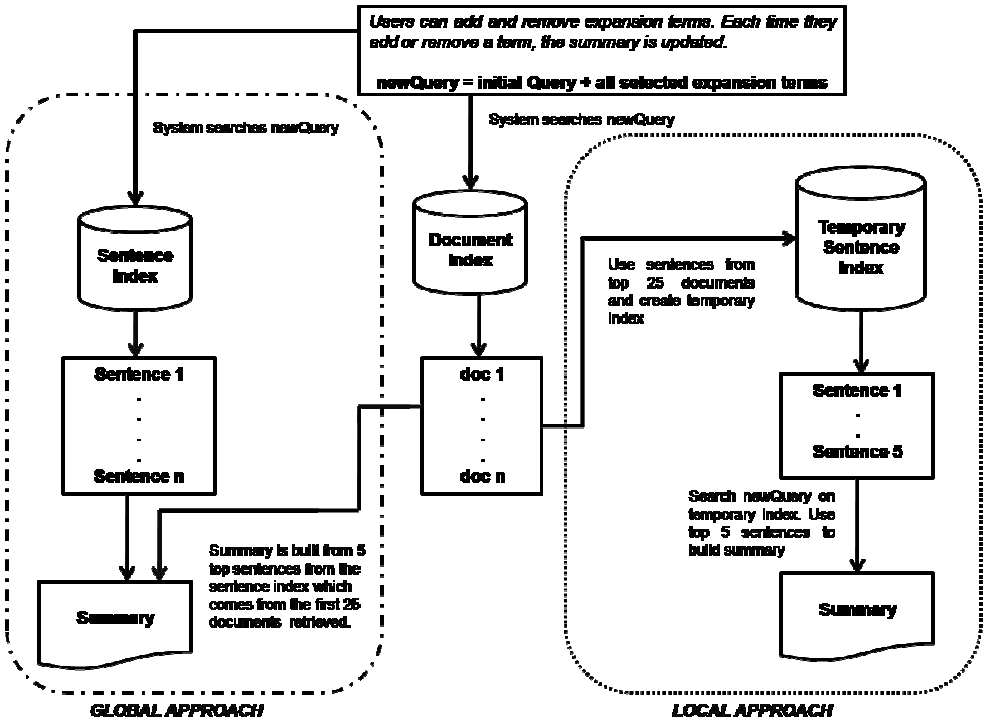


Figure 6: Graphical Representation of local and global approach in generating summary

In the global approach, the new query was searched again the sentence index itself and the top 5 sentences which came from the top 25 documents in *NewQueryDocs* were used to build the summary. However, as the system was using statistics, such as term frequencies and inverse document frequencies, from the complete document corpus, we failed to see in some cases the relevance between the sentences in the summary and the documents retrieved when the search was performed.

For the local approach, we used the top 25 documents in *NewQueryDocs* to extract their respective sentences from the sentence index. We created a temporary index with the sentences in order to be able to make use of the frequencies and statistic local to the sentences from the 25 documents. The new query was then searched against the temporary

index and the top 5 sentences used to construct the summary. This approach took slightly longer to compute the summary, but we felt that the quality of sentences in the summary justified the time. We found that there was a more obvious and intuitive relationship between the sentences in the summary and the documents that were retrieved, had the new query been searched for. In Figure 7, we compare summaries generated by the two approaches for the same query. In Figure 8, we show an example of a summary generated on the interface.

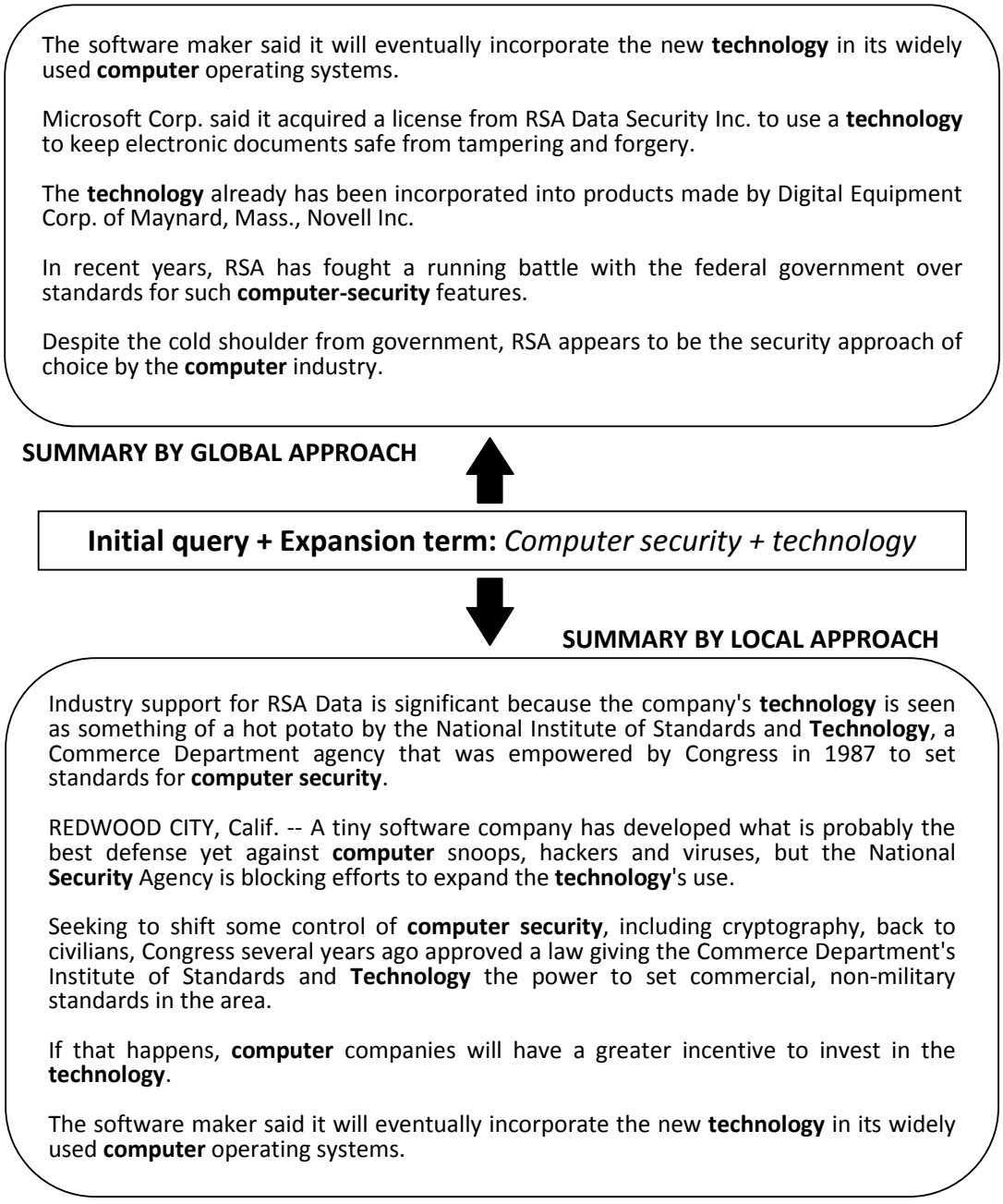


Figure 7: Comparison of local and global summaries

If the initial query was *'banking'* and the searcher select *'investment'* as one possible expansion word, the new query is *'banking investment'*. Thus the summary will show the top five sentences from the top twenty five documents that would be retrieved if the user searched this new query. This example is shown in Figure 8 below.

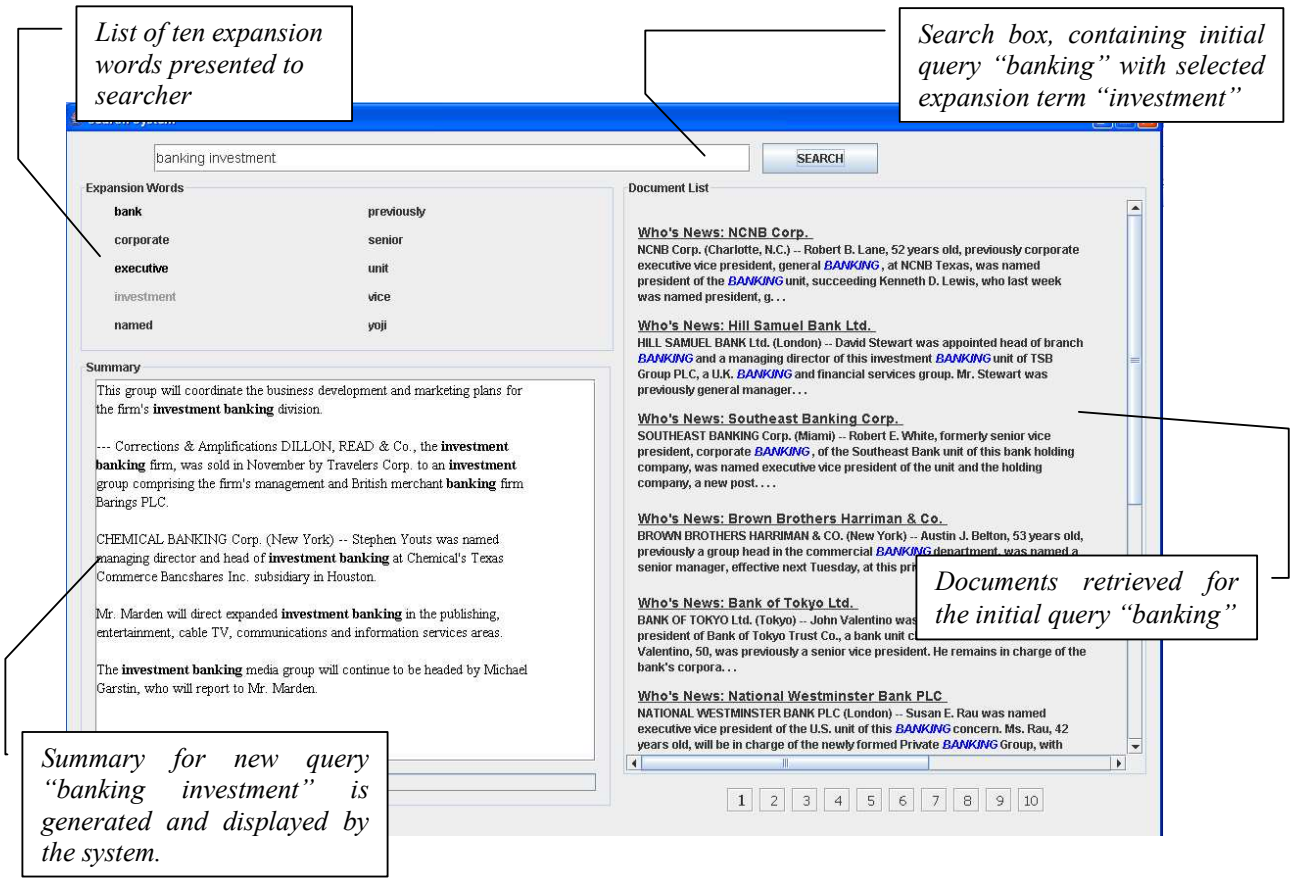


Figure 8: Screen shot of system with example search

Chapter 4

COMPARISON OF THE EFFECTIVENESS OF THE EXPANSION ALGORITHMS

As mentioned in section 3.4, the two expansion algorithms being considered in this project are the Rocchio feedback algorithm and the *wpq* weighting algorithm. In this section, the performance of the two algorithms will be compared in order to determine which one will be used in the IR system of this project.

4.1 TREC Topics and Relevance Assessments

TREC-5 (Voorhees and Harman, 1995) query topics were used in this study to compare between the two expansion algorithms. The TREC topics are made up of different sections as shown in Figure 9. However, for the purpose of this study, the title from each topic was used as the initial query because it is more similar to what users would type.

```
<top>
<num> Number: 251
<title> Exportation of Industry

<desc> Description:
Documents will report the exportation of some part of U.S. Industry to
another country.

<narr> Narrative:
Relevant documents will identify the type of industry being exported,
the country to which it is exported; and as well will reveal the
number of jobs lost as a result of that exportation.
```

Figure 9: Example of query topic from TREC-5

For each of the topic, there is a list of relevant documents that has been constructed manually by TREC assessors. Each document in the test collection is assessment as being relevant (1) or not relevant (0) to the topic. A segment of the relevance assessment file is shown in Figure 10. This allowed us to calculate the number of relevant documents for each query topic from TREC.

```
251 0 AP880215-0135 0
251 0 AP880223-0252 0
251 0 AP880224-0148 0
251 0 AP880224-0153 1
251 0 AP880224-0180 0
251 0 AP880224-0296 0
251 0 AP880224-0303 1
251 0 AP880225-0284 1
251 0 AP880226-0027 0
251 0 AP880228-0078 0
```

Figure 10: Fragment of relevance

4.2 Alternative versions of the *wpq* algorithm

As mentioned in section 3.4.2, we implemented different versions of the *wpq* algorithm in order to work around the problem of having to weight and rank a large number of terms. Therefore, we implemented and compared the following three versions of the algorithm:

The first version of the algorithm used the full text of the articles to generate expansion terms. This was found to introduce an unacceptable delay in our system and even if it did not affect the main aim of this study, we felt that it would have negative impact on searchers during the user experiments.

Secondly, we implemented a version of the algorithm using only the leading paragraph of each document, instead of the full text of the document. Leading paragraphs in the news articles are assumed to introduce the topic area of the article. Therefore, they are important because they provide a brief overview of what the article will deal with. This approach was justified by the results reported in (White et al., 2007), where they used document titles or text snippets, rather than the full text of documents, to generate expansion terms using pseudo relevance feedback. White et al. found a higher degree of similarity between those expansion terms and the terms generated from query logs that capture searcher's query refinement behaviour. Unfortunately, this approach deteriorated the performance of the algorithm despite reducing the time taken to generate expansion terms.

Finally, we adopted the approach of ranking all extracted terms by their *tf-idf* score. We then weighted only the top 50 terms using the algorithm. This value was chosen because searchers are only shown the top 10 expansion terms. Even after terms appearing in the initial query would be removed, there would be enough alternatives to display to the user. This version of the algorithm proved to be efficient and showed high performance. The results of the comparison between Rocchio and the different versions of *wpq* are reported in section 4.4.

4.3 Methodology

For each topic, we generate 10 expansion terms using both the Rocchio and *wpq* algorithms. Since searchers can choose any combinations of expansion words, there are 1023 possible combinations of terms for one search. Every combination of terms is added to the original query to form a new query. The new query was searched and a set of documents was retrieved. We adopted a full-freezing approach, where documents used to generate the list of expansion terms were not ranked among the retrieved documents. This approach is common standard when the performance of query expansion techniques based on relevance feedback is assessed (Chang et al., 1971).

The number of relevant documents found in the top 25 retrieved documents for the new query was recorded and averaged for each topic. The highest number of relevant documents retrieved for each topic was also recorded. The recorded information for all topics was averaged to get a mean score and a maximum score for each of the algorithms. We performed statistical testing on the data set using a single factor ANOVA[7].

[7] Analysis Of Variance test with level of significance at 5%

4.4 Findings

According to the ANOVA test carried out, there is no significant difference between the performances of the Rocchio algorithm and the different versions of the *wpq* algorithm. Table 3 shows the average and maximum score with no expansion is used, Rocchio algorithm and different versions of wpq algorithms. Average score refers to the average number of relevant documents retrieved in the top 25 documents, according to the TREC relevance assessment described in section 4.1. Maximum score is the highest number of relevant documents that were retrieved on average for the TREC topics in the top 25 documents.

	Average Score	**Maximum Score**
No Expansion (Initial Query)	1.64	1.64
Rocchio	1.92	3.84
***wpq* (Full Text)**	1.97	3.84
***wpq* (Leading Paragraph)**	1.51	2.66
***wpq* (Top 50 *tf-idf* terms)**	1.96	3.84

Table 3: Performance of Rocchio and different versions of wpq algorithm

4.5 Discussion

The *wpq* algorithm performs better than the Rocchio algorithm when the full text is used to generate expansion terms. This might be because our implementation of the *wpq* algorithm factors the fact that high frequency terms improve recall and very infrequent terms improve precision. On the other hand, the Rocchio algorithm under study only reweights terms using constant α and β, where set α is the weight for terms in relevant documents and β is the weight for non relevant documents. α and β are set to 1.0 and 0.5 respectively, without consideration that some terms will be more useful than others.

When ranking extracted terms by their *tf-idf* score before being weighted, the *wpq* algorithm shows a similar performance for the maximum score at 3.84. A very slight decrease in performance of the average score is noted, from 1.97 with the full text to 1.96. However, we feel that the reduction in time taken to generate terms justifies the almost negligible decrease in performance.

Using leading paragraphs, in our study, did improve the efficiency of the algorithm in terms of time, but deteriorated its performance. In fact, the average score for the *wpq* algorithm with leading paragraphs was worse than the score for the initial query. We attribute this performance to the fact that Wall Street Journal articles have a very unstable structure, in that the leading paragraph does not necessarily contain the most important terms that are present in the full text of the document.

Following these comparisons and the results obtained, in our search system, we decided to use the implementation of the *wpq* algorithm which weights 50 terms with the highest *tf-idf* score.

Chapter 5

USER STUDY: DESIGN

A laboratory-based user study was carried out for this project to evaluate whether additional information in the form of query biased summaries support searchers in completing their search tasks. User experiments were undertaken because it was important to capture the search behaviour of different types of searchers given the interactive nature of the system. The aim of the experiments is to study how searchers complete their search tasks using different systems which provide different levels of support. In this chapter, different aspects of the experiment will be described.

5.1 Systems

For this project, three search systems were implemented and used in the experiments: non-QE system which did not provide any expansion feature, IQE system with the possibility of adding terms interactively by the users and SummaryIQE system where users were given additional information of how their choice of expansion words would affect their queries. The three systems are explained in the following sections.

5.1.1 Non-QE System

In this system, users submitted their queries and the system would list the relevant documents retrieved with the document considered most relevant ranked highest in the list. No form of query expansion was available to the searcher and the documents retrieved were displayed 10 per page with their title and the two first sentences. The full text of documents was displayed in a separate window on request. This system resembles standard search systems and was a strong baseline system because searchers are very familiar with such a search environment. A screenshot of this system is included in Appendix A.

5.1.2 IQE System

The IQE system provided a query expansion feature where subjects could choose the terms to reformulate their query from a list generated by the system. Users submitted their queries to the system and the list of documents retrieved would be displayed as described in 5.1.1. Additionally, the system assumes the top 10 documents retrieved are relevant and uses them to generate a list of possible expansion words. Terms are extracted from those documents, stop words are removed and the rest of the terms are weighted according to the *wpq* algorithm (described in 4.5.2). The 10 terms with the highest score are displayed to the user as possible choices. Searchers can reformulate their initial query by adding or removing terms. Reformulated queries can be submitted to the system again. A screenshot of this system is included in Appendix A.

5.1.3 SummaryIQE System

The SummaryIQE system provides an interactive query expansion feature as mentioned in 5.1.2. However, an additional functionality in this system is to provide users with additional information as to how their choice of expansion words affects the results that would be

retrieved. Therefore, when searchers click on an expansion word, the system adds this word to the initial query and retrieved documents from the document-index that match the query. Sentences from the top 25 documents in that document set are indexed to create a temporary sentence index.

The new query is searched against the temporary index and the top 5 sentences retrieved are used to compile a query biased summary as described in section 3.5 where a screenshot of the system can also be found. The summary is displayed to the searcher in an attempt to show the kind of documents that would be retrieved if they choose to add the term to their query. The same procedure takes place when users decide to remove a term from their query.

5.2 Tasks

The tasks that searchers had to perform were derived from TREC-5 topics (Voorhees and Harman, 1995) and required subjects to gather information about the topic area in a maximum of 15 minutes. The tasks were classified in three categories depending on their hardness as calculated in TREC-5 (Voorhees and Harman, 1995). Table 4 shows the hardness measure of the three tasks and the complete description of all tasks are given in Appendix B.

Topic	Title	Hardness Measure[8]
288	Weight Control/Diet	0.3770
258	Computer Security	0.2184
300	Air Traffic Control Systems	0.1032

Table 4: Topics used in the user experiment

5.3 Subjects

A total of 18 subjects took part in the experiment and demographic information was gathered through the questionnaire before the experiment. Table 5 summarises the demographic information about the subjects.

Age	23.4 years (20 to 33)
Gender	13 Male, 5 Female
Online Search Experience	8.8 years
Use of suggested words	Yes – 50% , No – 50%
Computer Use frequency	Daily
Search frequency	Daily
Background	14 Computer Science, 4 Non-Computer Science

Table 5: Demographic information about subjects

[8] Lowest score = hardest
TREC-5 (Voorhees and Harman, 1995)

Subjects who use suggested terms in their online searches do so mainly for misspellings or grammatical mistakes. They find suggested terms useful in some situations, for example, when they do not get any results for their initial query, or when they are unsure about the keyword to submit to the system.

5.4 Experimental Design

This section describes how the user study was design. It explains our approach to collecting data from the experiments. The task and system allocation is also discussed.

5.4.1 Data Collection

Subjects completed a total of 8 questionnaires (included in the CD that accompanies this report) each at different stages in the experiment. There was a pre-experiment questionnaire, which gathered demographic information about subjects. Through this questionnaire, we also investigated subjects' search experience and regular use of computers and online searching. Before each task, subjects were asked to fill a short questionnaire which asked them to rate the task they were about to carry out. Semantic differentials were used in this questionnaire to gather subjects' familiarity with the topic (Unfamiliar – Familiar), the perceived difficulty (Difficult – Easy) and whether they found it interesting or not (Not interesting – Interesting). A questionnaire after each task included questions with semantic differentials as well as 5-point Likert scales to gather data about subjects' search experience with the system and the interface. After the three tasks, subjects completed a final questionnaire, which required them to rate the three systems and the tasks.

During the task, users were also asked to note the title of any documents they thought as useful for their task. There were also open ended questions for general comments on search systems and interfaces. Finally, there was an informal interview with subjects.

To gather information about subjects' interaction within each system, system logging was used. The logs recorded information about the queries being searched and also contained the expansion terms the searcher adds or removes. If searchers view the complete text of a news article, the unique identifier for that document, the document number, would also be recorded in the logs. In Figure 11, the fragment of a log file is shown.

```
<message>Searching...computer security hack</message>
<message>Added expansion word-codes</message>
<message>Searching...computer security hack codes</message>
<message>Document viewed position:1 Document ID  AP881216-0065</message>
<message>Document viewed position:8 Document ID  AP881105-0083</message>
<message>Page Viewed: 2</message>
<message>Added expansion word-fraud</message>
<message>Searching...computer security hack  fraud</message>
<message>Document viewed position:1 Document ID  AP880422-0139</message>
<message>Searching...computer security fraud</message>
<message>Document viewed position:1 Document ID WSJ900507-0106
<message>Added expansion word-access</message>
<message>Searching...computer security fraud access</message>
```

Figure 11: Fragment of log file

5.4.2 Task and System Allocation

Task and system were allocated using a Greco-Latin design. In this design approach, we rotate the order in which users attempt the tasks and also the systems they use to perform the task so that the order of task and system is not repeated in one block of users. Table 6 shows the task and system allocation for a block of 9 participants. S_1, S_2 and S_3 refer to *Non-QE System*, *IQE System* and *SummaryIQE System* respectively whereas T_1, T_2 and T_3 represent easy, medium and hard tasks.

	System (S) and Task (T) Allocation		
U_1	S_1T_1	S_2T_2	S_3T_3
U_2	$S_1\ T_2$	$S_2\ T_3$	S_3T_1
U_3	$S_1\ T_3$	$S_2\ T_1$	S_3T_2
U_4	$S_2\ T_1$	S_1T_2	S_3T_3
U_5	$S_2\ T_2$	S_1T_3	S_3T_1
U_6	$S_2\ T_3$	S_1T_1	S_3T_2
U_7	$S_3\ T_1$	S_1T_2	$S_2\ T_3$
U_8	$S_3\ T_2$	S_1T_3	$S_2\ T_1$
U_9	S_3T_3	S_1T_1	$S_2\ T_2$

Table 6: Task and system allocation for 9 searchers

5.5 Experimental Protocol

1. Subjects are given a brief overview of the experiment and asked to sign consent form.
2. Subjects are asked to complete a pre-experiment questionnaire to capture background information about their search experience.
3. For each of the three systems:
    a) Subjects are shown their task
    b) Subjects are asked to fill in a short questionnaire about their perceived difficulty of the task they were about to perform
    c) Subjects are allowed up to15 minutes to complete their task.
    d) After completion of the task, subjects are asked to complete a questionnaire about their search experience and the task.
4. After completing three tasks using the three different systems, subjects are asked to answer a post-experiment questionnaire to compare their different search experiences with the different systems.
5. Debriefing; subjects are asked for qualitative feedback in an informal interview and their comments are recorded.

5.5 Objectives of the user study

We carried out the user study to evaluate whether our proposed system, the *SummaryIQE System*, supports searchers effectively in their search tasks. We are particularly interested in the support provided by the query biased summaries. Therefore, we focus on three main points.

First, we will investigate whether summaries are effective in helping searchers complete their information seeking tasks. We consider the number of relevant documents retrieved for each task and searchers' overall search experience indicative of the effectiveness of summaries.

Secondly, we will evaluate the uptake of IQE. We are particularly interested in studying whether summaries lead to an increase in the use of suggested terms to reformulate initial queries. We will use the data from the system logs to answer this question.

Finally, we will focus on hard and unfamiliar tasks as we believe that it is in those cases that searchers are in most need of support from IR systems. Therefore, we will study how searchers perform their most difficult and their unfamiliar tasks by considering the relevant documents they retrieve for these tasks and the number of times they reformulate their initial queries using suggested terms.

Chapter 6

USER STUDY: RESULTS AND DISCUSSION

In this section, we report the findings from the user experiments carried out. We address three main questions in presenting our results. Firstly, we study whether summaries are effective at helping searchers complete their search tasks. Next, we address the question of whether additional information in the form of summaries increases the uptake of IQE. Finally we focus on the support provided to searchers for hard and unfamiliar tasks. We also report other findings, where we discuss patterns observed during the experiments.

6.1 Are summaries effective at helping searchers complete their information seeking tasks?

In this section, we address one of the main aims of this study as presented in section 1.1. We use data from the questionnaires, system logs and the informal interviews to present our findings.

6.1.1 Number of relevant documents found by searchers

The two systems with query expansion outperform the baseline system in terms of relevant documents retrieved by searchers during their tasks. Searchers on average found more relevant documents on the systems that provide an expansion feature, with or without the summary. The number of relevant documents for each subject was computed from the list of documents that searchers had on each of their task sheets, that is, the title of the news articles that they felt were relevant to their search tasks.

Relevant Documents	***Mean***	***Standard Deviation***
Non-QE System	4.67	1.46
IQE System	5.78	2.05
SummaryIQE System	5.61	2.33

Table 7: Relevant documents for the three systems

Table 7 shows that users retrieved more relevant documents on the *IQE System* and *SummaryIQE System*. We attribute this to the fact that searchers had additional support from the system. Suggested expansion terms helped users to reformulate their queries and by adding terms suggested by the system, searchers found more documents relevant to their search tasks. 77.8% of users felt that they were most supported in their tasks by the systems with an IQE feature. One way repeated measures ANOVA test revealed no significant statistical difference between the three systems at $F(2,51) = 1.65$, *ns*[9], but subjects found slightly more documents with the *IQE System* than they did with the *SummaryIQE System.*

[9]No significance at $p < 0.05$

One likely explanation for this might be that subjects took more time to perform their tasks when using *SummaryIQE System.* 10 out of 18 subjects took the maximum time allocated when using *SummaryIQE System,* while only 4 and 5 users spent 15 minutes on their tasks with *Non-QE System* and *IQE System* respectively. This argument is reinforced by the fact that subjects reported that the time allocated on the *SummaryIQE System* was the least sufficient among the three systems as shown in the Table 8. The fact that users had to spend longer time on their selection of terms by reading the summary might explain the need for extra time. Also, since searchers felt that the time was not sufficient implies that they did not have enough time to view as many documents as they would have liked. Therefore, the number of relevant documents that they retrieved with the *SummaryIQE System* was slightly less than those they retrieved using the *IQE System*.

	Non-QE System	***IQE System***	***SummaryIQE System***
Sufficient time allocated *(On a 5 point Likert Scale)*[10]	4.1	4.3	3.9

Table 8: Subjects' views on time allocation for tasks

Despite finding slightly less relevant documents with the SummaryIQE System, searchers were most satisfied with the information they found with the SummaryIQE System. Table 9 shows searchers' satisfaction with the information found. In the after-task questionnaire, searchers were also asked whether they found enough information for the task and the results are included in Table 9.

(On a 5 point Likert Scale)	***Non-QE System***	***IQE System***	***SummaryIQE System***
Found enough information for the task	3.44	3.78	3.83
Satisfied with information found	3.28	3.39	3.89

Table 9: Subjects' views on information found for their tasks

6.1.2 Preferred search system

In this section, we present findings gathered from the post experiment questionnaire and the informal interviews about subjects' overall impression of the systems they used. In Figure 12, we report the overall system preference and Table 10 shows subjects' choice for the search system that was easiest to use and provided most support.

[10] Unless otherwise stated, all 5 point Likert scale has the following range:
Strongly disagree (1) to Strongly agree(5)

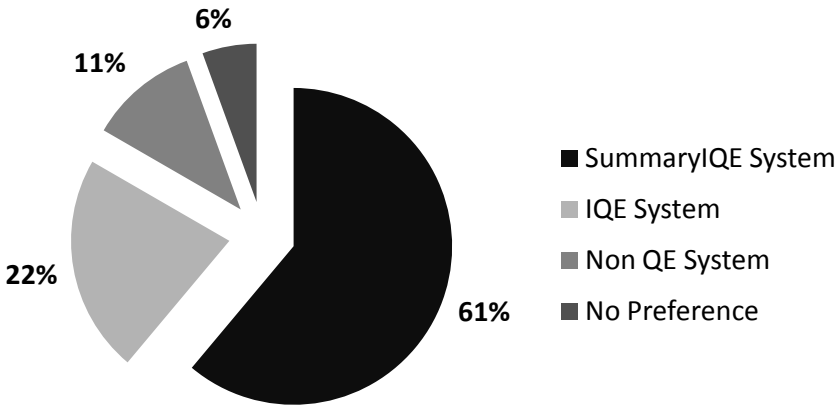


Figure 12: Overall system preference

	Non-QE System	***IQE System***	***SummaryIQE System***	***No Preference***
System that provided most support for tasks	11.11%	27.78%	50.00%	11.11%
System easiest to use	22.22%	22.22%	33.34%	22.22%

Table 10: Subjects' views on search systems

During the informal interview, 10 out of 18 searchers commented on the usefulness of the summaries in providing them with an idea of how their results will be impacted depending on different combinations of expansion terms. Subjects in general felt that if they could see how expansion terms appear in documents, it helped them make a decision about the usefulness of suggested terms in their respective search tasks. They compared this feature to the *IQE System*, where no summaries were provided, as they had to add the expansion terms to their queries, search for the new query to find out what happens to their results. Some users reported on finding this frustrating and time consuming.

For the *IQE System* and the *SummaryIQE System,* subjects were asked in the questionnaire about the suggested terms; whether the suggested terms were useful in completing the task, whether they understood how to use the terms and whether they found them intuitive. A 5-point Likert scale from 1 *(Strongly disagree)* to 5 *(Strongly agree)* was used for the user responses. Table 11 shows a summary of their response including the mean (M) and standard deviation (SD) for each system (the higher the mean, the better).

	IQE System		***SummaryIQE System***	
	M	SD	M	SD
Suggested terms were useful in completing tasks	3.2	1.0	3.8	0.8
Terms were intuitive	3.3	0.7	3.9	0.8
Understood how to use suggested terms	3.7	0.8	4.2	0.7

Table 11: Subjects' views about suggested terms

A single factor ANOVA test reveals significant statistical difference between the two systems [$F(1,34) = 6.57$, $p = 0.028$] for the intuitiveness of suggested terms. Searchers also thought that suggested terms were more useful in completing their tasks with the *SummaryIQE System*. A single factor ANOVA test shows significant statistical difference between *IQE System* and *SummaryIQE System* [$F(1, 34) = 4.69$, $p = 0.037$]. Subjects found terms on the system with the summary more intuitive and useful because they could see the context in which the terms appeared and could compare this to their own information need. If the term appeared in a context similar to what they were looking for, they could use the terms to reformulate their query.

There was no significant statistical difference in how subjects understood how to use suggested terms with single factor ANOVA test. However, during the informal interviews after each experiment, a number of searchers commented on not knowing how to use the expansion terms. They claim a simple listing of the terms provided them with alternative search terms but they were not supported to use it. They could not, for example, see the benefits or drawbacks of adding more than one term at a time. If they added several terms to their query, performed a search and did not like the results, they could not know which of the words they added led to the results not matching their information needs. This is the reason why many participants felt that they could be more specific in their queries with the *SummaryIQE* system.

6.2 Do summaries increase the uptake of IQE?

In this section, we present findings about the number of query reformulations with each of the three systems. We also study the extent to which searchers use suggested expansion terms are to reformulate their queries.

6.2.1 Query reformulation[11]

We studied the subjects' interaction with each of the search systems through the logs and determined the number of times a query was reformulated. Table 12 shows searchers tend to reformulate their queries a higher number of times using the *IQE* system (approximately 7 reformulations per user) although there was no significant statistical difference with single factor ANOVA test at $F(2,51) = 0.525$, *ns*.

[11] *'Query reformulation'* in the rest of this report refers to subjects submitting a new query to the system after the initial query. The new query may include suggested expansion terms or may consist only of searcher's own choice of terms.

	Non-QE	***IQE***	***SummaryIQE***
Average number of reformulations	5.89	7.44	5.89

Table 12: Average number of reformulations with the three systems

We believe this is so because, if a searcher thinks one of the suggested terms might be useful for their query, they will have to add that term to their initial query and search for that new query, to be able to find out the impact of adding that term, on their results. This explains why participants performed a higher number of query reformulations using the *IQE* system.

On the other hand, searchers reformulate their initial queries less on *SummaryIQE* because when they add a suggested term to their query, they do not have to search for that 'new query' to know what impact it will have on their results. This is so because support is provided to them in the form of summaries. When searchers add expansion term or terms, the summaries show them a snippet of what their results would contain if they decide to submit the query. Therefore, if they do not like the impact on the results, they can remove one or all of the expansion terms. This is not the case with the *IQE System* where searchers only find out what happens to the results after reformulating their search. Therefore, the decrease in number of reformulations with the *SummaryIQE System* is justified.

In addition, the percentage of query reformulations that include suggested expansion terms is shown in Figure 13. 46.23% of query reformulations on the *SummaryIQE System* include one or more suggested expansion terms compared to *IQE* where only 30.59% of the reformulations include expansion terms. This shows that users of the *IQE* system are reformulating their query more manually, by adding terms they pick up from reading the full text of documents or from their task description. This implies that they are not sufficiently supported by the IQE feature in the system. With the *SummaryIQE System*, subjects used the suggested terms more often to reformulate their queries because before the reformulation they are aware of how their results will be affected through the summaries.

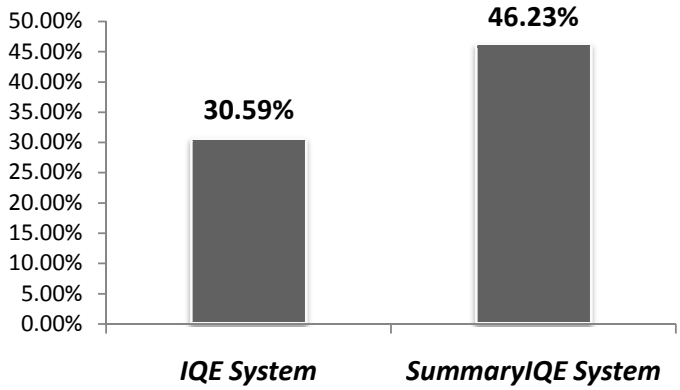


Figure 13: Percentage of query reformulations that include suggested expansion terms

6.3 Do summaries support searchers in completing hard or unfamiliar tasks?

This question directly relates to the objectives of the user study. We analyse whether summaries provide additional support to searchers when they are faced with unfamiliar topics or difficult tasks.

6.3.1 Task difficulty

In the post experiment questionnaire, searchers were asked to rate the tasks they had performed in terms of their difficulty (*1 – Easiest to 3 – Most Difficult*). From the informal interview, we found out that depending on the user's background, they identified different tasks as being the most difficult. For example, a number of male subjects found the task about *'Weight Control/Diets'* most difficult because they were not familiar with the topic, while searchers without a computer science background thought the *'Computer Security'* task was the hardest. The interviews showed that there was a correlation between unfamiliarity of the task and its perceived difficulty.

We identified the hardest task for each user and studied their performance related to the system they searched the task on. Table 13 shows the number of documents that searchers found relevant for their hardest task and the number of query reformulation on each system.

	No-QE System	**IQE System**	**SummaryIQE System**
Number of relevant documents	3.2	4.7	6.0
Number of query reformulations	11.3	5.8	7.4

Table 13: Average number of relevant documents and query reformulations for hard tasks on each system

Single factor ANOVA revealed no statistical difference at $F(2,13) = 3.03$, *ns,* between the means. However, for their hardest tasks, searchers used suggested terms more for query reformulations with the SummaryIQE System than they did with the IQE System. Figure 14 shows the percentage of query reformulations that included one or more suggested expansion terms.

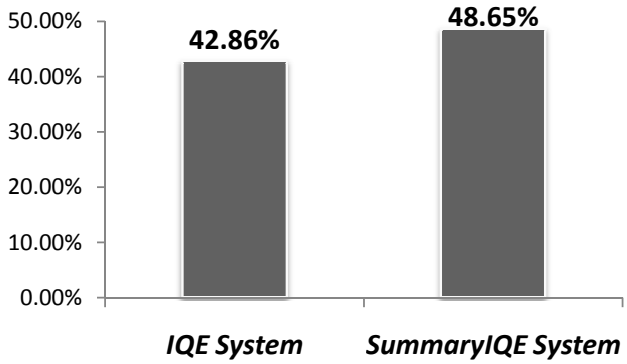


Figure 14: Percentage of query reformulations that include suggested terms for hard tasks

We also used the before-task questionnaire to identify tasks perceived as unfamiliar by searchers. The question about task familiarity was posed using a 5-point Likert scale [Unfamiliar (1) to Familiar (5)]. We considered 1 and 2 to represent unfamiliarity with the topic area and identified searchers' unfamiliar tasks and the systems they performed these tasks on. We present the results in Table 14, including the averages for the number of relevant documents and the number of query reformulations for each system.

	No-QE System	**IQE System**	**SummaryIQE System**
Number of relevant documents	4.6	6.5	5.2
Number of query reformulations	6.6	10.0[12]	5.5

Table 14: Average number of relevant documents and query reformulations for hard tasks

Therefore, searchers tend to reformulate their queries most on the *IQE System* for unfamiliar tasks. However, a single factor ANOVA test found no significant statistical testing between the three systems. We also studied the percentage of query reformulations for unfamiliar tasks that included suggested terms and we present the data in Figure 15.

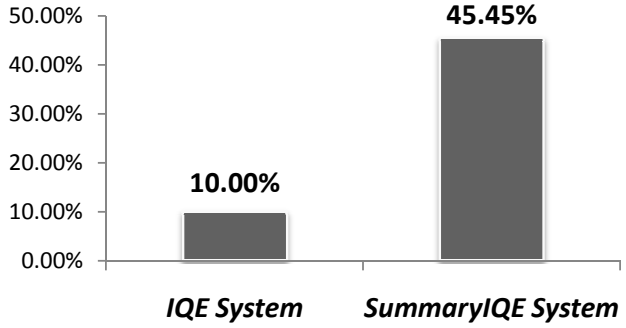


Figure 15: Percentage of query reformulations including suggested terms for unfamiliar tasks

Figure 14 and 15 show the percentage of query reformulations that included suggested terms with the two systems for hard tasks and unfamiliar tasks respectively. In both cases, we observe a higher uptake of IQE with the *SummaryIQE System*. 48.65% of query reformulations for hard tasks with the *SummaryIQE System* included one or more suggested terms compared to 42.86% with *IQE System*. The difference is more acute for unfamiliar tasks where searchers included suggested terms in 45.45% of their query reformulations with *SummaryIQE System* compared to only 10.00% with the *IQE System*.

These findings are in line with the data gathered during informal interviews at the end of user experiments despite no significant statistical difference with a single factor ANOVA test. In fact, a number of searchers said that they were more likely to use IQE when the tasks were hard or unfamiliar. It appears that the summaries in the SummaryIQE System led to an increase in the use of suggested terms to reformulate queries. We attribute the

[12] This result for *IQE System* is affected by an outlier of 29 query reformulations for one user.

higher percentage for the use of IQE for the unfamiliar tasks to the fact that in an unknown environment, as described in section 1.1, searchers are more likely to try different terms that have been suggested to them by the system because they do have prior knowledge of the search area. On the other hand, searchers might think a task is difficult because they are unable to find documents relevant to the search topic and not because they are unfamiliar with the topic itself.

6.4 Other Findings

In this section, we report various other results from the experiments that are not directly related to the effectiveness of summaries.

6.4.1 Search Experience

After each task, subjects were asked to rate their search experience. We present the findings in Table 15. On the Likert scale, the value of 5 is the best value, therefore, in Table 13, the higher the mean, the better.

On a 5 point Likert Scale	***Non-QE System***		***IQE System***		***SummaryIQE System***	
	M	**SD**	**M**	**SD**	**M**	**SD**
Frustrating - Pleasing	3.1	0.9	3.3	0.9	3.4	0.8
Stressful - Relaxing	3.4	0.8	3.4	0.8	3.5	0.7
Not intensely absorbing- Intensely Absorbing	3.2	1.0	3.4	0.8	3.1	1.0
Boring- Interesting	3.1	1.2	3.3	1.0	3.5	1.1

Table 15: Search experience for each system (M –Mean; SD – Standard Deviation)

Single factor ANOVA shows no statistical significance between the means for the three systems implying no system provided users with a significantly different search experience. Some searchers, however, commented that for tasks that they were not familiar with, the *Non-QE System* was often frustrating because they did not know in which direction to take the search to get documents that met their information needs.

6.4.2 System usability

A number of participants liked the simplistic design of the interface and commented that it allowed them to carry out their search tasks easily. As expected, 55% of users in the experiment thought the first system, *Non-QE*, was the easiest to learn to use. This is because most users are familiar with the layout which is used by most online search engines. 2 out of 18 subjects preferred the *Non-QE system* overall, because they liked the familiarity and were hesitant to learn how to use a new feature even if they were aware that it could help them retrieve better results for their tasks. Informal discussions with those users revealed that they believe that they can get to the relevant documents simply by browsing different pages of results. In fact, our logs of system interactions shows that with the *Non-QE* system, searchers tend to browse through more pages than they would with a system that provide an IQE feature.

Chapter 7

CONCLUSION

In this final chapter, we provide an overall discussion of our results and discuss the implications of the study that was carried out. We also include concluding comments on the work undertaken and outline the scope for future work.

7.1 Discussion and Implications

Despite having the potential to be effective, the uptake of IQE has been limited. Some researchers suggest that the lack of uptake is one of the reasons why IQE is not widely present in IR systems (Dennis et al., 1998). Informal discussions with subjects from our study show that searchers tend to use IQE more when they are not familiar with the task or the task is difficult or complex. This seems to be in line with previous studies (Fowkes and Beaulieu, 2000) which showed that searchers are more likely to use IQE when they cannot express their information need easily.

Given the limited use of IQE, White and Marchionini (White and Marchionini, 2000) presented an alternative way of displaying suggested terms. Their system, *RealTime*, involves presenting terms to users at the query formulation stage. However, they acknowledge that presenting to searchers expansion terms before they have seen any search results can lead to searchers adding 'erroneous' terms to their query. White and Marchionini attribute this problem to the fact that searchers cannot predict the effect of adding a term to their query and argue that this is a problem of '*functional visibility*', not just with their approach, but with all systems that implement a feature based on relevance information. The issue of *'functional visibility'* as indentified in (Beaulieu and Jones, 1998) is discussed in section 2.3 of this report.

Our study had identified two reasons for the limited uptake of IQE: searchers do not know where suggested terms come from and they are unsure about the impact of their choice of terms on the results. We tried to address the second issue by providing searchers with additional information in the form of query-biased summaries. The aim of the summaries was to enable users to predict the effect of adding query expansion terms to their initial query. Results from our laboratory-based experiments show the following:

- ✓ The majority of searchers (61%) preferred the *SummaryIQE System* overall as they understood how to use the expansion terms and felt they were intuitive.
- ✓ Most of the searchers felt that the *SummaryIQE System* supported them most in their search tasks (50% for *SummaryIQE System*, 28% for *IQE System* and 11% for *Non-QE System*).
- ✓ Searchers retrieved the most relevant documents for their hardest task on the *SummaryIQE System.*
- ✓ There was a 16% increase in the uptake of IQE with the *SummaryIQE System*.

- ✓ For unfamiliar tasks, 45% of query reformulation with our proposed system included suggested terms compared to 10% with the *IQE System*.
- ✓ Searchers reformulate their queries most frequently on the *IQE System*, but only 30% include suggested terms compared to *SummaryIQE System* where 46% of query reformulations include suggested terms.

Data gathered from informal interviews with searchers show that they are more likely to use IQE if they know the impact that their selection of expansion terms will have on their results. We observed a 16% increase in the use of expansion terms to reformulate queries and report that searchers feel more confident about using IQE when they are aware of the context that the suggested terms appear in. Our findings also show that in circumstances where users find the task difficult or unfamiliar, they are likely to find more relevant documents with the *SummaryIQE* system and use suggested terms more frequently to reformulate their queries. In fact, for unfamiliar tasks, 45% of query reformulations on the *SummaryIQE System* included one or more suggested terms compared to 10% with the *IQE System*. Even though our system was preferred overall by searchers, it is important to note that 50% of the participants in our experiments do not use suggested terms in their daily online searches. Therefore, this might have contributed to the reluctance of some users to use the query expansion feature.

The *Non-QE System* was preferred by some users because it was familiar and easiest to learn to use. The additional support for query expansion did not seem to affect the time taken by users to complete their task. In fact, users of *SummaryIQE* took the longest time to perform their tasks. This may be because searchers needed extra time to evaluate summaries before deciding on their search strategy.

7.2 Scope for future work

Following the user study, searchers provided some suggestions for the interface. We also include in this section, the scope for future work on the system itself.

7.2.1 Interface improvements

A number of participants suggested that sentences in the summary should be clickable, so that their source document can be directly accessible. The argument was that, at times summary sentences were direct answers to their query and it would save a lot of time if the document containing the sentence could be accessed directly. Other users commented that the summary should be constructed and displayed when they hover on the expansion terms instead of clicking on them. They claim that the option to click on the term should be reserved if they actually wanted to add the term to their initial query in order to retrieve documents that match the new query. This approach however might not work in the current system as summaries are generated not only for one term but also for a combination of selected expansion terms.

7.2.2 Future work on the system

One area for future work is to address the first issue relating to the limited uptake of IQE, which is the development of a feature to explain to users of IR systems, where the

suggested terms came from. This should overcome the issue of trust (Ruthven, 2003) that many users have with systems that suggest terms.

Other scope for future work would be exploring the nature of expansion terms whereby, phrases that relate to the area of search could also be suggested alongside single terms. This might help more experienced searchers in their tasks. Semantic relationships between suggested terms and initial query could also be displayed to searchers because it is often assumed that searchers can exploit semantic relationship between words. However, experiments (Blocks et al., 2000) have shown that searchers can fail to make the most of their understanding of semantics even if the system supports the recognition of semantic relationships. This problem should be investigated as despite humans having the intuitive ability to exploit semantic relationship, they need the system to explicitly provide them with evidence before they are able to make decisions.

In addition, further improvements can be made as far as the summaries are concerned. Clustering could be used so that summaries within the same topic are grouped together in an attempt to summarise the topic area for searchers. A similar clustering approach could also be implemented for expansion terms. Given that searchers shown only 10 terms, it might be useful to group similar terms together and suggest other related terms to searchers.

7.3 Conclusion

We present in this report a study that tried to address the problem of limited uptake of IQE. Our system, *SummaryIQE*, provides users with query biased summaries that are updated in real time depending on a searcher's choice of expansion terms. Summaries aim at providing searchers with additional information on how their choice of expansion terms affects the results retrieved by the system.

Our study shows strong evidence that summaries are effective in supporting searchers to complete their information seeking tasks. We reported an increase in the uptake of IQE when searchers were provided with additional information through the query biased summaries. We studied whether summaries supported searchers for their most difficult and unfamiliar tasks and our results showed that 45% of query reformulations for unfamiliar tasks include suggested expansion terms when summaries were provided compared to 10% in the absence of summaries. In addition, there were significant statistical differences between the *IQE System* and the *SummaryIQE System* when it comes to how intuitive users found suggested terms. This is due to the fact that users feel more supported when they can see the context in which suggested terms appear in documents that would be retrieved if they submitted their initial query with the term. Our system was preferred by 61% of searchers because of the additional support. We report therefore, that summaries appear to be effective in helping searchers to complete their tasks. The additional support provided by the query biased summaries was also important for difficult and unfamiliar tasks.

If the reasons for the limited uptake of IQE are overcome, its potential effectiveness could provide users of IR systems with enhanced support in formulating their queries. We view this study as a step towards this direction.

REFERENCES

Beaulieu, M. (1997) *Experiments with interfaces to support query expansion* Journal of Documentation, 53: 8-19.

Beaulieu, M and Jones, S. (1998) *Interactive searching and interface issues in the Okapi best match probabilistic retrieval system.* Interacting with Computers, 10(3): 237-248.

Blocks, D., Binding, C., Cunliffe, D., and Tudhope, D. (2002) *Qualitative evaluation of thesaurus-based retrieval.* Proceedings of the 6th European Conference in Digital Libraries. Rome. Lecture Notes in Computer Science 2458, pages 346-361.

Brajnik, G., Mizzaro, S. and Tasso, C. (2002) *Strategic Help in User Interface for Information Retrieval.* Journal of the American Society for Information Science and Technology, 53(5): 343 – 358.

Brandow, R., Mitze, K. And Rau, L.F. (1995) *Automatic condensation of electronic publications by sentence selection.* Information Processing and Management 31(5): 675-685.

Chang, Y. K., Cirillo, C. And Razon, J. (1971) *Evaluation of feedback retrieval using modified freezing, residual collection & test and control groups.* The SMART retrieval system – experiments in automatic document processing. Salton G., ed. pages 355- 370, Prentice-Hall.

Chirita, P.A., Firan C.S. and Nejdl W. (2007) *Personalised query expansion for the web.* Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval, Amsterdam, pages 7-14.

Dengel, A. (2006) *Six thousand words about multi-perspective personal document management.* Proceedings IEEE-EDM: IEEE Internat. Workshop on the Electronic Document Management in an Enterprise Computing Environment, Key Note Paper, pages 1-10.

Dennis, S., McArthur, R., & Bnjza, P.D. (1998). *Searching the World Wide Web Made Easy? The cognitive load imposed by query refinement mechanisms.* Proceedings of the third Australian document computing symposium (ADCS'98), Sydney, pages 65-71.

Edmundson, H.P. (1964) *Problems in automatic abstracting.* Communications of the ACM, 7(4): 259 – 263.

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.

Fowkes, H., & Beaulieu, M. (2000) *Interactive searching behaviour: Okapi experiment for TREC-8.* Proceedings of the IRSG 2000 colloquium on IR research, Cambridge, UK.

Gospodnetic, O. and Hatcher, E. (2005) *Lucene in Action* Manning Publications, USA.

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

Koenemann, J. and Belkin, N.J. (1996) *A case for interaction: A study of interactive information retrieval behaviour and effectiveness.* Proceedings of CHI 96 International Conference on Human Computer Interaction, Vancouver, Canada, pages 205-212.

Harman, D. (1988) *Towards Interactive Query Expansion*. Proceedings of the 11th annual international ACM SIGIR conference on research and development in information retrieval, Grenoble, pages 321-331.

Koenemann, J. and Belkin, N.J. (1996) *A case for interaction: a study of interactive information retrieval behaviour and effectiveness*. Proceedings of the Human Factors in Computing Systems Conference (CHI'96), Zurich, pages 205-212.

Magennis, M and van Rijsbergen, C.J. (1997) *The potential and actual effectiveness of interactive query expansion.* Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval. Philadelphia, pages 324 – 332.

Manning D. Christopher, Raghavan Prabhakar and Schutze Hinrich. (2008) *Introduction to Information Retrieval.* Cambridge University Press 2008.

Marchionini, G. (1989) *Information-seeking strategies of novices using a fulltext electronic encyclopedia.* Journal of the American Society for Information Science, 40(1): 54 – 66.

Robertson, S.E. (1990) *On term selection for query expansion.* Journal of Documentation, 46(4): 359-364.

Rocchio J. (1971) *Relevance Feedback in Information Retrieval. In: The SMART Retrieval System: Experiments in Automatic Document Processing.* Salton G., ed. pages 313- 323, Prentice-Hall.

Ruthven, I. (2003) *Re-examining the potential effectiveness of interactive query expansion.* Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Toronto, pages 213-220.

Ruthven, I. Tombros, A. and Jose, J.M. (2001) *A study on the use of summaries and summary-based query expansion for a Question-answering task.* Proceedings of the 23rd BCS European Annual Colloquium on Information Retrieval Research (ECIR 2001), Germany, pages 41-53.

Salton,G. and McGill, M. (1983) *Introduction to Modern Information Retrieval* McGraw-Hill, New York

Spink, Amanda, Bernard J. Jansen, and H. Cenk Ozmultu. 2000. *Use of query reformulation and relevance feedback by Excite users.* Internet Research: Electronic Networking Applications and Policy, 10(4):317–328.

Sullivan, M., Borgman, C.L. and Wippern, D. (1990) *End-users, mediated searches and front-end assistance programs on dialog: A comparison of learning, performance and satisfaction.* Journal of the American Society for Information Science, 41(1): 27 – 42.

Tombros, A and Sanderson, M. (1998) *The advantages of Query Biased Summaries in Information Retrieval.* Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, pages 2-10.

Teevan, J, Dumais, S.T. and Horvitz,E. (2005) *Personalizing search via automated analysis of interests and activities.* Proceedings of the 28th annual international ACM SIGIR conference on Research and Development in Information Retrieval, Brazil, pages 449-456.

Van Rijsbergen, C.J. (1979) *Information Retrieval*, 2nd Edition, Butterworth, London

Van Rijsbergen, C.J. (1995) *Information retrieval and informative reasoning,* Lecture Notes in Computer Science, Springer, Berlin, pages 549 – 559

Vollaro, A.J. and Hawkins, D.T. (1986) *End-user searching in a large library network: A case study of patent attorneys.* Online, 10(4): 67 – 72.

Voorhees, E.M. (1994) *Query Expansion using lexical-semantic relations.* Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval, Dublin, pages 61 – 69.

Voorhees, E.M. and Harman, D. (1995) *Overview of the Fifth Text REtrieval Conference (TREC-5).* In Proceedings of Text Retrieval (TREC-5), National Institute of Standards and Technology, Gaithersburg, MD 20899, USA.

White, R.W. and Ruthven, I. (2006) *A study of interface support mechanism for interactive information retrieval.* Journal of the American Society for Information Science and Technology, 57 (7): 933-948.

White, R.W., Clarke, C.L.A. and Cucerzan, S. (2007) *Comparing Query Logs and Pseudo-Relevance Feedback for Web-Search Query Refinement.* Proceedings of the 30th annual ACM SIGIR conference on Research and development in information retrieval, The Netherlands, pages 831 – 832

White, R.W. and Marchionini, G. (2006) *Examining the effectiveness of real time query expansion.* Journal of Information Processing and Management, 43(3): 685-704.

White, R.W. Ruthven, I. and Jose, J.M. (2005) *A study of factors affecting the utility of Implicit Relevance Feedback.* Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, Brazil, pages 35 – 42.

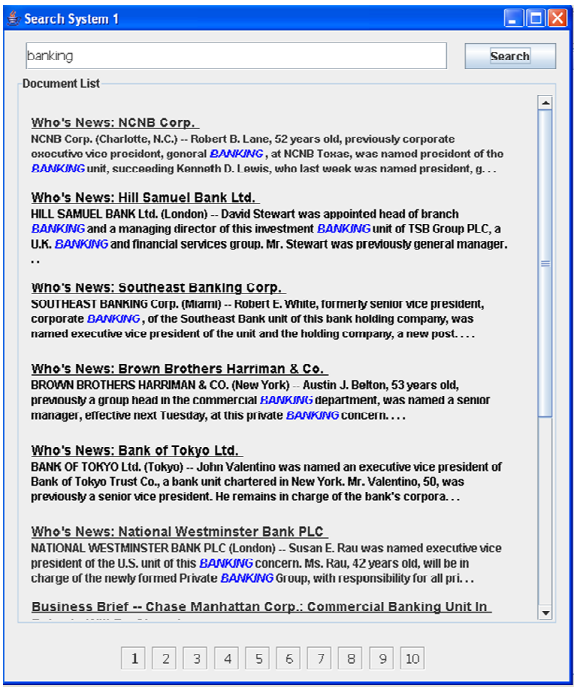
Xu, J and Croft, W.B. (2000) *Improving the effectiveness of information retrieval with local context analysis.* ACM Transactions on Information Systems, 18(1): 79 – 112.

Xu, J and Croft, W.B. (1996) *Query expansion using local and global document analysis.* Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval, Switzerland, pages 18 – 22.

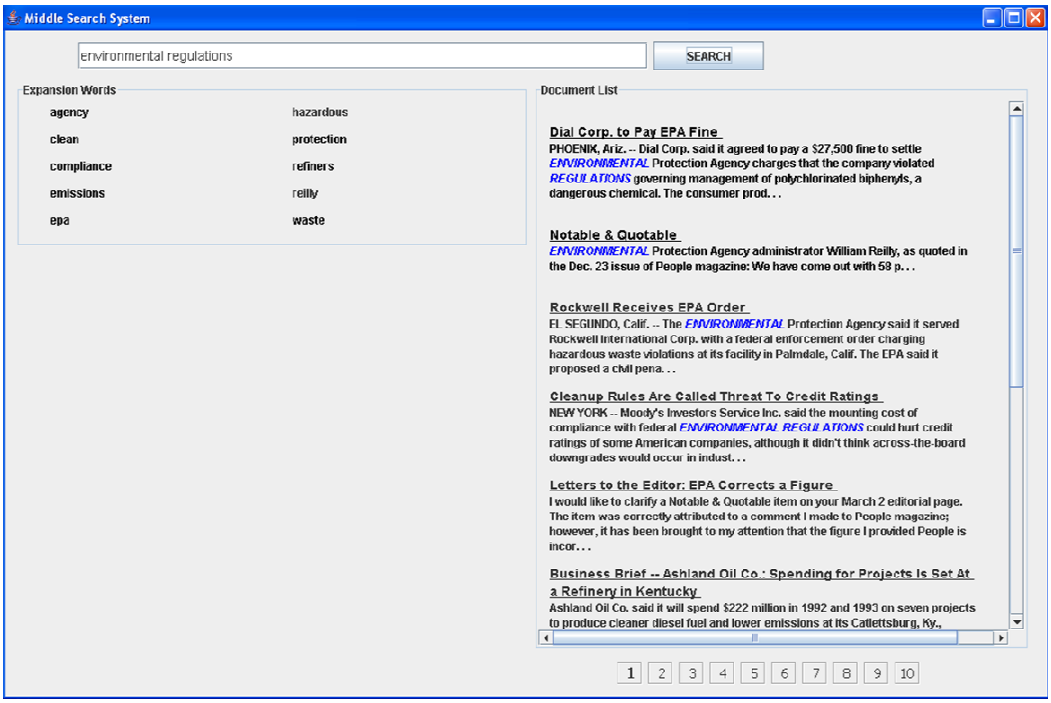
Smart, J.F. *Integrate advanced search functionalities into your apps*, 25/09/2006, http://www.javaworld.com/javaworld/jw-09-2006/jw-0925-lucene.html 21/01/2009

APPENDIX A

- *Non-QE System*



- *IQE System*



APPENDIX B

- **Easy Task**

Weight Control/Diets

Sports nutrition leaders say the key to controlling weight, while fuelling yourself properly for activity, is to eat the right kind of foods at the optimum time. More than a third of U.S. women and nearly a quarter of U.S. men are trying to lose weight at any one time, according to a report by the NIH.

Evidence presented at an NIH Conference showed that diets fail the majority of time. Some therapists believe that overweight people need to learn to accept themselves, forget their obsession of losing weight and get on with their lives.

Useful documents will report on weight control and diets in the U.S.

- **Medium Task**

Computer Security

Identify instances of illegal entry into sensitive computer networks by non-authorized personnel.

Illegal entry into sensitive computer networks is a serious and potentially menacing problem. Both 'hackers' and foreign agents have been known to acquire unauthorized entry into various networks. Items relative this subject would include but not be limited to instances of illegally entering networks containing information of a sensitive nature to specific countries, such as defence or technology information, international banking, etc.

Items of a personal nature (e.g. credit card fraud, changing of college test scores) should not be considered relevant.

- **Hard Task**

Air Traffic Control Systems

The U.S. National Air Traffic Controllers Association frequently warns that air traffic control systems around the world are inadequate and that personnel are frequently overworked. Is there evidence in investigation data that the mechanical and human systems worldwide frequently contribute to airliner casualties?

To be useful, an item will concern a commercial airline accident anywhere in the world and consider, if not outright attribute, the casualty to a difficulty (either mechanical or human) with an air traffic control system.

APPENDIX C: Source Code Components

The source code consists of the following packages:

[1] Backend

[2] Interface

[3] Relevance

[4] Opennlp.util.tools

[5] SentenceSplit

[1] [2] [3] were developed as part of the project

[4] [5] are open source natural language processing tools available at http://opennlp.sourceforge.net

The search system for this project was developed using the Lucene API available at http://www.lucene.apache.org