Formulating Complex Queries Using Templates

by

Hao Zhang

A thesis

presented to the University of Waterloo

in fulfillment of the

thesis requirement for the degree of

Master of Applied Science

in

Management Sciences

Waterloo, Ontario, Canada, 2008

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

While many users have relatively general information needs, users who are familiar with a certain topic may have more specific or complex information needs. Such users already have some knowledge of a subject and its concepts, and they need to find information on a specific aspect of a certain entity, such as its cause, effect, and relationships between entities. To successfully resolve this kind of complex information needs, in our study, we investigated the effectiveness of topic-independent query templates as a tool for assisting users in articulating their information needs. A set of query templates, which were written in the form of fill-in-the-blanks was designed to represent general semantic relationships between concepts, such as cause-effect and problem-solution. To conduct the research, we designed a control interface with a single query textbox and an experimental interface with the query templates. A user study was performed with 30 users. Okapi information retrieval system was used to retrieve documents in response to the users' queries.

The analysis in this paper indicates that while users found the template-based query formulation less easy to use, the queries written using templates performed better than the queries written using the control interface with one query textbox. Our analysis of a group of users and some specific topics demonstrates that the experimental interface tended to help users create more detailed search queries and the users were able to think about different aspects of their complex information needs and fill in many templates.

In the future, an interesting research direction would be to tune the templates, adapting them to users' specific query requests and avoiding showing non-relevant templates to users by automatically selecting related templates from a larger set of templates.

Acknowledgements

First, I would like to take this opportunity to express the deepest appreciation to my supervisor Professor Olga Vechtomova for her invaluable support, encouragement, supervision and useful suggestions throughout this research work. Without her guidance and persistent help this thesis would not have been possible.

I would like to thank Professor Frank Safayeni and Professor Rob Duimering. They spent time in reading this paper and made valuable comments on it.

In addition, I would like to thank all faculty, staff, and graduate students in the department of Management Sciences for their warm-hearted support during my study at the University of Waterloo.

Finally, I would especially thank my family for their love and support throughout my life.

Table of Contents

List of Tables

List of Figures

Chapter 1 Introduction

Nowadays, many users engage in information-seeking behavior normally because they want to resolve some problems, or achieve some goals, for which their current state of knowledge is inadequate. Among these users, some of them have relatively simple or general information needs, but some other users who are familiar with a certain topic may have more specific or complex information requirements [43]. For those users who already have some knowledge of a subject and concepts, finding information on some specific aspects of a certain entity, such as its cause, effect, relationships between entities, instances, solutions to combat/reduce/alleviate problems, statistical data, etc. is the main target when they are involved in the information retrieval activities [1]. Such information needs have been referred to as complex needs.

In order to successfully resolve complex information needs, an IR (Information Retrieval) system, should firstly encourage and help users articulate the characteristics of potentially useful information objects based on their current knowledge, and secondly, to enhance performance it should have a suitable retrieval model to handle the rich information provided by the users. In this study, we focus on the first part of which an IR system can help users formulate their complex information needs by expressing specific semantic relations between the entities of their interest.

Growing interest in interactive systems for answering complex questions led to the development of the Complex Interactive Question Answering (CIQA) task, which was introduced for the first time at Text REtrieval Conference (TREC) 2006. The purpose of the interactive aspect of CIQA is to provide a framework for participants to investigate interaction in the question answering (QA) context and an opportunity for other researchers to become involved in QA [3]. Thereafter, many researchers have successfully done some studies in this area. Most interfaces to experimental and operational IR systems are developed based on the assumption that users often have a difficult time articulating their information needs and typically pose very short queries which usually are between two and three words in length [6]. This gap between an information need and a query statement can significantly affect the effectiveness of the IR system.

Kelly et al. [2] conducted an experiment within the framework of the High Accuracy Retrieval from Documents (HARD) track of TREC 2004. In this experiment, to obtain additional information a feedback form was sent to users. In this form, Kelly asked the user to supply information in response to three prompts/questions which were followed by large textboxes to encourage users to talk more about his/her information need. The study indicates that the IR system performance could be improved when users respond to the prompts. Furthermore, the results demonstrate a strong relationship between query length and performance, but the form does little to help users articulate the relationships between concepts and entities.

In our study, we propose a novel approach to query formulation, whereby the system suggests to the user a list of templates in the form of fill-in-blanks questions. In each template, a topic-independent semantic relation, such as “cause-effect” or “problem-solution”, is presented. An example of a template is “What effect does ____ have on ____?”. In section 3, a brief review of RST (Rhetorical Structure Theory) is provided and some related works are discussed, which are the construction foundation of our query templates. Then based on the manual analysis of the descriptions and narratives of 100 TREC topics (#301-400) from TREC 6 and 7 ad-hoc tracks, we conducted a set of possible query templates.

Okapi retrieval system is based on Robertson and Spark Jones’ Probabilistic Model with user relevance feedback [8]. In our experiments, by employing this experimental IR system with the goal of improving retrieval performance, we implemented two query formulation interfaces: a template-based interface and a single-textbox interface. It is reasonable to expect that a template–based query formulation interface using semantic relations in document ranking may be more effective than the single-textbox interface. To explore the effectiveness of user queries formulation, a user study with thirty participants was conducted, in which the experiment was designed using the Latin Square principle. Section 4 reports the experiment design and protocol.

To reach the major goal of our user study, which is to investigate whether system users can articulate complex information needs by filling in topic-independent templates, we explored two research questions:

> RQ1: Are users able to articulate complex information needs using topic-independent query template?

> RQ2: When used with a bag-of words retrieval system, do the queries formed using template-based interface lead to better search results than queries formed using a single-textbox interface?

Our user study results demonstrate that even though the templates based query formulation is not as easy to use as single-textbox interface, the retrieval performance of our a-bag-of-words system has been improved by using these templates. The user study results are discussed in section 5 in detail based on the analysis of the retrieval performance, search queries, and user satisfaction.

In section 6 and 7, some search topics used in our experiments and several related user activities are analyzed. Finally our study is concluded and some future research directions are outlined.

Chapter 2 User Information Needs and Interactive IR Systems

2.1 User Information Needs

Users may have relatively simple or general information needs when they are not familiar with a subject or have limited knowledge of its concepts. Thus, when they interact with IR systems, their information needs are likely to be relatively simple, and also because of their insufficient knowledge of a topic, it is normally difficult or impossible for users to express their information needs precisely. Belkin et al. [4] in their well-known ASK (Anomalous State of Knowledge) hypothesis stated that “an information need arises from a recognized anomaly in the user's state of knowledge concerning some topic or situation and that, in general, the user is unable to specify precisely what is needed to resolve that anomaly”.

Consider the following scenario: an environment policy consultant, who is working on a project to investigate the effectiveness of a government environment policy, needs to write a report to discuss a government environment policy’s effectiveness for a long term view. He needs to find information about the local environment changes after the policy been executed, and also he needs to explain the reasons why the policy caused some worse situations. Moreover, he would like present his advice to improve it. Because of his specific and complex information requirements, he is not interested in those documents explaining the environment protection policy generally. Instead, the documents which could answer his information needs as we showed above will be the best results to help him.

In contrast, an example scenario is that a first-year university student, who starts to work on a course paper of environmental issues which are related to government policy, has no knowledge of this subject and does not have a clear idea what she/he needs to know. Thus, some basic information about the subject would be the searching start point.

In general, users with a high level of familiarity with a certain subject area may have very specific or complex information needs [1]. In other words, by having some knowledge of a subject and its concepts or entities, users may want to find information on a specific aspect of a certain entity, such as its cause, effect, how it can be prevented, what can be done to achieve it, or a relationship between entities, rather than the general introduction of concepts. Thus in order to successfully resolve complex information needs, an IR system, should be effective in encouraging and helping users to articulate queries by choosing the correct words to represent their information problems.

It is well understood and documented that the gap between an information need and a query statement can significantly affect the effectiveness of the IR system, and increased query length leads to increased system performance as measured by recall and precision [2]. There are many IR systems developed based on the assumption that users often have a difficult time articulating their information needs and typically pose very short queries which usually are between two and three words in length [6]. While it might be the case that the difficulty that users have with articulating their information needs causes their queries to be short, it has been argued that one reason users pose short queries is because traditional search interfaces encourage them to do so [5]. Thus, there is an apparent mismatch between what interfaces encourage users to do, what users are doing and what has been demonstrated to result in good retrieval.

In this paper, as an investigation tool, the topic-independent query templates are constructed to assist users in articulating their complex information needs.

2.2 IR Models

Baeza-Yates and Riberio-Neto [7] stated that one central problem of IR system is the issue of predicting which documents are relevant and which are not. Such a decision is usually dependent on a ranking algorithm which is the core of IR systems. The IR model is used to represent the query and text indexing, and calculate the relevance of retrieved documents. A ranking algorithm operates according to basic premises regarding the notion of document relevance and distinct sets of premises yield distinct IR models.

There are two basic concepts behind the IR models: index term and weight [7]. In many systems, an index term is simply any word which appears in the text of a document in the collection. Therefore, index terms are used to represent the document contents. On the other side, given a set of index terms for a document, not all terms are equally useful for describing the document contents. To define the varying relevance discriminating ability which distinct index terms have, the assignment of numerical weights to each index term of a document is invoked.

Based on the mathematical methods used to construct the models, the IR models can be defined with set theoretic models, algebraic models, and probabilistic models. In the set theoretic models, documents and queries are represented as sets of index terms and the basic three set theoretic models include the Boolean model, fuzzy set model, and Extended Boolean model. In the algebraic models, documents and queries are represented as vectors in a t-dimensional space; the vector space model and latent semantic indexing model are the popular ones. In the probabilistic model, the framework for modeling document and query representations is based on probability theory. In this study, the three classic models in IR area, Boolean, Vector, and Probabilistic, are discussed below.

2.2.1. Boolean Model

The Boolean model is a simple retrieval model based on set theory and Boolean algebra. It predicts that each document is either relevant or non-relevant by considering that index terms are present or absent in a document.

Chiaramella and Chevallet (1992) [9] defined that in the Boolean model, index terms of the document d will be linked by the operator “and”, and a query q is composed of index terms linked by three operators: not, and, or. As a result, the index term weights are assumed to be all binary, i.e., $w_{ij} \in \{0,1\}$. Boolean expressions specify the queries by having precise semantics. Thus, a query which is essentially a Boolean expression can be represented as a disjunctive or conjunctive vector (i.e., in disjunctive normal form--DNF).

A query q is a conventional Boolean expression. Let $\overrightarrow{q_{dnf}}$ be the disjunctive normal form for the query q. Further, let $\overrightarrow{q_{cc}}$ be any of the conjunctive components of $\overrightarrow{q_{dnf}}$. The similarity of a document d_j to the query q is defined as formula 2.1

$$sim(d_j, q) = \begin{cases} 1 & if \;\; \exists \overrightarrow{\mathrm{q_{cc}}} \,\big|\, (\overrightarrow{\mathrm{qcc}} \in \overrightarrow{q_{dnf}}) \wedge (\forall k_i,\, \mathrm{g_i}(\overrightarrow{\mathrm{d_j}}) = g_i(\overrightarrow{\mathrm{q_{cc}}})) \\ 0 & \text{otherwise} \end{cases} \qquad (2.1)$$

If $\mathrm{sim}(d_j,q)=1$ then the Boolean model predicts that the document d_j is relevant to the query q (it might not be). Otherwise, the prediction is that the document is not relevant.

The main advantages of the Boolean model are the clean formalism behind the model and its simplicity. Actually, many search engines used commercially, including the Google Advanced Search [10] still include some concepts of Boolean model, but returning a ranked set of documents.

However, the Boolean model suffers from few disadvantages. First of all, the retrieval strategy is based on a binary criterion without any notion of a grading scale, which may prevent good retrieval performance. Partial relevance cannot be obtained through Boolean models since exact matching mechanism has only two results, match and not match. Secondly, it is not easy to structure a query expression to respond the real information need. Mostly, it is difficult and awkward to express queries in terms of Boolean expressions. Also, since Boolean searching is an exact term matching process, which matches index terms in queries and those in documents, it is not capable of coping with such situation when none of the query terms are present in the document, while according to the user's information needs, the document is relevant.

Despite all of these factors, the Boolean model is still used in many IR systems due to its simplicity of framework and easy implementation. In particular, most of the specialized search engines in the legal domain, such as Westlaw, are based on the Boolean model.

2.2.2 Vector space model

The vector space model is an algebraic model representing natural language documents in a formal manner through the use of vectors (such as index terms) in a multi-dimensional linear space. This model proposes a framework in which partial matching is possible by assigning non-binary weights to index terms in queries and documents.

There are three important factors to understand in the vector space model: the calculation of the degree of similarity, the choice of terms, and the calculation of weights. In the vector space model, documents are represented as t-dimensional vectors of index terms (keywords). The vector space model proposes to evaluate the degree of similarity of the document d_j with regard to the query q as the correlation between the vectors $\vec{d_j}$ *and* $\vec{q}$.

Relevance rankings of documents in a keyword search can be calculated using the assumptions of document similarities theory, by comparing the degrees of angles between each document vector and the original query vector. Instead of calculating the actual angles between vectors, the cosine of the angle between vectors is calculated and compared. A cosine value of zero means that the query and document vector were orthogonal and had no match (i.e. the query term did not exist in the document being considered).

It is a simple way to represent a document by its full set of words. In this case, the retrieval system adopts a full text indexing which could be handled with modern computers. However, in fact, it is better to reduce the set of representative keywords by the elimination of stopwords and the use of stemming, i.e. reducing multiple word-forms to one word-form (e.g., “walking”, “walked”, “walks” and “walker” are transformed to “walk”).

In the vector space model, the similarity is qualified by measuring the raw frequency of a term k_i inside a document d_j. Such term frequency is usually referred to as the tf factor and provides one measure of how well that term describes the document contents. Furthermore, by measuring the inverse of the frequency of a term k_i among the documents in the collection, which is referred to as inverse document frequency (idf), this factor is given by

$$idf_i = \log \frac{N}{n_i} \qquad (2.2)$$

where N is the total number of documents in the system and n_i is the number of documents in which the index term k_i appears.

Hence, the best known term-weighting schemes use weights which are given by tf*idf.

The classic vector space model had both local and global parameters incorporated in the term weight equation (known as the tf-idf):

$$\omega_{i,j} = f_{i,j} \cdot \log \frac{N}{n_i} \qquad (2.3)$$

The main advantages of the vector space model are its simplicity, its good retrieval performance, and its ranking formula. Its partial matching strategy which is simple and fast allows retrieval of documents that approximate the query conditions. Also, its term-weighting scheme and cosine ranking formula improves retrieval performance compared to the Boolean models. It yields ranked answer sets which sort the documents according to their degree of similarity to the query. However, the vector space model has some disadvantages also. The index terms are assumed to be mutually independent. However, there are some relations between words in a language. Furthermore, the answer sets have no query expansion or relevance feedback within the framework of the vector model.

2.2.3 Probabilistic model

The probabilistic models were proposed and extensively researched by many researchers and various experiments were conducted, evaluated and analyzed. In this section, a description of Robertson and Sparck Jones' Probabilistic Model is discussed in detail.

Robertson and Sparck Jones developed their initial probabilistic model based on a theory of relevance weighting [11]. Robertson and Walker incorporated variables, such as length of documents, term frequency within documents and query, into their model and Sparck Jones et al. further improved the framework of the probabilistic model [12].

In their probabilistic models, the fundamental idea is as follows. Given a user query, there is a set of documents which contains exactly the relevant documents and no other, as the ideal answer set. At query time, the properties of an ideal answer set are not known and users only know that there are index terms whose semantics should be used to characterize these properties. Thus an effort has to be made at initially guessing what they could be and then this initial guess generates a preliminary probabilistic description of the ideal answer set. The system will repeatedly use the information from an interaction with the user after he/she judges the first top documents' relevance to refine the description of the ideal answer set.

The central elements in probabilistic models are terms which are used to describe queries and documents. Their attributes, such as term frequency, term presence or absence in a document, can be used to predict the probability of document relevance. Furthermore, the probabilistic model treats retrieval as a ranking process which ranks the documents in the collection in the order of their probability of relevance. If term i is present in a document, it has the weight

$$W_i = \log\frac{P_i(1-\overline{P_i})}{\overline{P_i}(1-P_i)}, \qquad (2.4)$$

where P_i is the probability of presence of term i in the relevant document set, and $\overline{P_i}$ is the probability of term i in the non-relevant document set [12]. The relevance weight of a document is the sum of the weights of all terms in the document.

For a more refined interpretation of the probabilistic model, the term incidence contingency table is introduced by Robertson and Sparck Jones [11], which is shown in table 2.1.

Table 2.1 The Term Incidence Contingency table

	Relevant	**Non-Relevant**	**Total**
Containing the term	r	n-r	N
Not containing the term	R-r	N-n-R+r	N-n
Total	R	N-R	N

where R is the number of relevant documents for this query, r is the number of relevant documents containing the term, n is the number of documents which contain the index term, and N is the number of documents in the collection. If we have the above relevance information, the term presence weighting function can be further developed to formula 2.5 as follows:

$$w = \log \frac{r(N-n-R+r)}{(R-r)(n-r)} \tag{2.5}$$

Estimation considerations give rise to a simple modification of the formula, namely to add 0.5 to all the central cells. Then we can derive a specific term relevance weighting formula 2.6:

$$RW = \log \frac{(r+0.5)(N-n-R+r+0.5)}{(R-r+0.5)(n-r+0.5)} \tag{2.6}$$

This formula gives relatively higher weight to query terms that have a high relevant document incidence and low additional non-relevant document incidence. It is well-behaved in extreme cases [12].

The probabilistic model of information retrieval system provides a good and comprehensible basis for a systematic exposition of the components of the formula and their interrelations. Also the model is useful to translate directly into a retrieval mechanism based on a simple query-document matching or scoring function. Furthermore, this model is well-founded on the mathematical theories and it is clear and substantial to design an IR system.

Okapi BM25 is a ranking function based on the probabilistic retrieval framework. The name of the actual ranking function is BM25 and it is usually referred to as "Okapi BM25", since the Okapi information retrieval system, implemented at London's City University in the 1980s and 1990s, was the first system to implement this function [11].

“BM25 is a bag-of-words retrieval function that ranks a set of documents based on the query terms appearing in each document, regardless of the inter-relationship between the query terms within a document (e.g., their relative proximity)” [12]. It is not a single function, but actually a whole family of scoring functions, with slightly different components and parameters. One of the most prominent instantiations of the function is as follows.

Given a query Q, containing keywords $q_1,...,q_n$, the BM25 score of a document D is:

$$\text{score}(D,Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i,D) \cdot (k_1+1)}{f(q_i,D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{\text{avgdl}})}, \qquad (2.7)$$

where $f(q_i,D)$ is q_i's term frequency in the document D, | D | is the length of the document D in words, and avgdl is the average document length in the text collection from which documents are drawn [12]. k1 and b are free parameters, usually chosen as k1 = 1.2 and b = 0.75. IDF(q_i) is the IDF (inverse document frequency) weight of the query term qi. It is usually computed as:

$$\text{IDF}(q_i) = \log \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5}, \qquad (2.8)$$

where N is the total number of documents in the collection, and $n(q_i)$ is the number of documents containing q_i [12].

When some relevance judgments are available, i.e. after the user has seen and judged several retrieved documents as relevant, IDF in Equation 2.7 is substituted with Relevance Weight (RW) as given in Equation 2.6.

2.3 Evaluation Methods

There are different reasons why evaluation of IR is important. For instance, providers of information resources require information about usage of these resources by users, and organizations working on improving search engines need effective methods for evaluating changes made to algorithms and user interface. Thus, the purpose of evaluation is to lead to improvements in the information retrieval process. On the other hand, goals of evaluation usually depend on the research interests of scholars. Some researchers may argue that the main goal of evaluation is to evaluate the power of indexing and search methodologies; however, some other foci of IR evaluation research are cognitive processes of users, human–computer interface, and characteristics of databases [13].

IR evaluation methods basically are divided into two broad categories: evaluation of performance and evaluation of outcome. Measures of performance are descriptive of what happens during the use of an IR system. Measures of outcome are descriptive of the results obtained [14]. In a system designed for providing information retrieval, the answer of the question “how well does the system work” basically includes two parts: the time and space efficiency; effectiveness of the results [7].

2.3.1 Recall and Precision

In response to a query, an IR system searches its document collection and returns an ordered list of responses, which is called the retrieved set or ranked list. The system employs a search strategy or algorithm to measure the quality of a ranked list. Usually, a better search strategy yields a better ranked list and also the better ranked lists help the user fill their information need.

With respect to a given query, the documents can be partitioned into four sets: relevant or not, retrieved or not; user says Yes/No, system says Yes/No. Precision measures the fraction of retrieved documents that are relevant while recall measures the fraction of all relevant documents that are retrieved [7]. To be more specific: precision = r/n and recall = r/R, where r = the number of relevant retrieved, R = total number of relevant documents, n = the number of retrieved documents. In other words, precision measures the efficiency of the search, while recall measures the breadth of the search. For example, a query Q has 100 related documents, and the IR system retrieved 200 documents in which there are 80 documents truly related the query topic. The Recall equals 80/100=0.8 and the precision equals 80/200=0.4. The result shows that the recall is higher but the precision is lower.

Both recall and precision are important factors to evaluate IR systems, but there is trade-off between recall and precision. For instance, the system can increase recall by retrieving more and this can decrease precision. In different user models, trade-off has a different meaning: a web searcher is going to look at the top 20 documents and a patent attorney wants all the relevant documents in the collection. The precision-recall curve which maps precision-recall pairs on a scatter plot shows the retrieval performance at each point in the ranking. Each query generates a separate P-R curve. The overall performance of the system is judged by merging these curves.

2.3.2 P@N and MAP

There are various approaches to using precision and recall as measures of evaluation. The most common strategy is to measure precision at fixed levels of recall for each query and build the average P-R curve by using the average precision at each recall level. Therefore, search methods are often compared on the basis of their average precision at a given level of recall. Hence, the average precision over recall levels and P-R curves are typical ways of evaluating IR method performance.

An alternative evaluation strategy is to compute precision and/or recall while holding the documents constant for each query. Both average precision and average recall can be computed by this method. Van Rijsbergen took this a step further and defined his E-measure, which is a weighted combination of precision and recall [7].

The first approach measures the precision at the same fixed levels of recall for all queries and averages the results, which are then presented on a curve. The second approach examines a fixed number of documents for each query and uses this information to compute a precision or recall score. The evaluation measure is then averaged over all the queries.

Precision at N documents retrieved (Precision@N) sets precision at fixed points in the ranking by counting the number of relevant documents in the top N documents in the ranked list returned for a topic. Since most users care about the results shown on the first 1 or 2 pages, P@10 or P@20 is normally used. However, this measure has a very large margin of error [15].

R-precision is defined as the precision after R relevant documents for the given topic are retrieved. Average precision is defined as average of precision at each relevant document retrieved and the precision of a non-retrieved relevant document equals 0. This measure is useful for evaluating the effectiveness of a ranked list for a single query. MAP is based on much more information than other measures [7]. It calculates the mean of the precision scores obtained after each relevant document is retrieved.

Besides all of these methods, different IR applications or tasks take totally different evaluation factors into account. For instance, for evaluating a Question Answering system, the rank of the retrieved documents is the main factor.

2.3.3 Statistical Tests

An evaluation study is not complete without some measurement of the significance of the differences between retrieval methods. Statistical significance tests provide this measurement.

The statistical tests can be extremely useful to evaluate the IR system performance because they provide information about whether observed differences in evaluation scores are really meaningful or simply due to chance. It is impossible to make such a distinction for an individual query, but when evaluation measures are averaged over a number of queries, one can obtain an estimate of the error associated with that measure and significance tests become applicable [13].

While most of the standard tests require normality and continuity assumptions that are unlikely to be exactly true for IR data, it is still possible to test the assumptions using simple diagnostic plots [16]. If they are indeed violated, there are a number of nonparametric tests which are reasonable alternatives.

The preliminary assumption or null hypothesis H0 will be that all the retrieval methods being tested are equivalent in terms of performance. The significance test will attempt to disprove this hypothesis by determining a p-value, a measurement of the probability that the observed difference could have occurred by chance. Prior to the experiment, a significance level α is chosen, and if the p-value is less than α, one can conclude that the search methods are significantly different. A smaller value of α (which is often set to 0.05) implies a more conservative test. Alternatively, the p-value can merely be viewed as an estimate of the likelihood that two methods are different. All significance tests make the assumption that the queries are independent. However, slight violations of this assumption are generally not a problem.

2.3.4 TREC

The Text REtrieval Conference (TREC), an annual event held by the US government's National Institute of Standards and Technology (NIST) in November, provides the infrastructure necessary for large-scale evaluation of text retrieval methodologies [17]. For each TREC, participants run their own retrieval systems based on the TREC collections, and the test sets of documents and questions offered by NIST. Their individual returned results, which are lists of the retrieved top-ranked documents, then will be judged and evaluated. Over the years, the number of participants has steadily increased and the types of tracks, which are the areas of focus, have greatly varied. In the fifteenth conference, TREC 2006, it had 107 participating groups from 17 different countries and seven tracks were explored [17].

The specifics of each TREC track are not relevant since the tracks are continuously modified. In TREC 2006 [17], five of the tracks ran in previous TRECs and explored tasks in question answering, detecting spam in an email stream, enterprise search, search on terabyte-scale document sets, and information access within the genomics domain. The two new tracks explored blog search and providing support for legal discovery of electronic documents. The common theme of all tracks is to establish an evaluation corpus to be used in evaluating search systems.

The conference participation procedures are as follows. Firstly, a call for participation is announced with defining the specifics of each task. After procuring documents and topics, each participant team conducts a set of experiments and then submits their results for judgment. In TREC, pooling, which is the process of selecting top-ranked documents obtained from multiple engines, merging and sorting them, and remaining the unique document identifiers, is used to determine the relevant documents. Relevance assessments are obtained, the submitted results are evaluated, and then at the annual meeting the findings are summarized and presented to all participants. After the meeting, all participants submit their summary papers.

Early TREC forums used data on the order of multiple gigabytes. Today, the types of data vary greatly, depending on the focus of the particular track. In TREC 2005, a terabyte data collection was proposed for one of the tracks. Thus, within a decade, the collection size has grown by three orders of magnitude from a couple of gigabytes to a terabyte. We will discuss the TREC collection data with documents and topics in details at chapter 3.

TREC, although successful, does have its shortcomings. As we discussed above, performance evaluation in retrieval systems involves both accuracy and performance assistance. TREC, however, only evaluates accuracy, paying little significance to processing times and storage overheads [7]. In terms of relevancy, common TREC criticism focuses on the mean of judging document-to-query relevancy. Moreover, although relatively effective, pooling does result in several false-negative document ratings which triggered out some new effectiveness measures to be proposed to avoid this problem [7]. In our study, we used TREC topics from No.301-450.

2.4 Okapi System

Okapi is the name given to a family of experimental retrieval systems that have been developed over the last three decades [18]. It is based on the Robertson/Sparck Jones probabilistic model of searching with relevance feedback [12], which was outlined in Section 2.2.3 of this thesis. Okapi started life as an online library catalogue system and was designed to use simple, robust techniques internally to present a user interface which requires no training and no knowledge of searching methods or techniques [18].

A typical Okapi system involves a search engine, indexing routines, interface system, and query model layer [19]. Basic Search System (BSS), as the search engine, provides efficient low level functionality for weighting and ranking searches; various interface systems provide different representations of the underlying BSS functionality; Query Model manages the interaction between the user interfaces and the BSS and supporting additional functions such as query expansion.

The Okapi search session involves following steps: firstly users enter search terms, and system pre-processes, parses and stems the search terms to convert the user input to the standard form used in indexing the documents in the databases; secondly, the system stems and weights the remaining terms, calculates the document score, and then presents a ranked list of matching documents to the user; thirdly, user provides relevance judgments feedback on the documents (called relevance feedback); finally, system selects terms from relevant documents, expands the query and performs a new search based on the expanded query, and then returns a new ranked list to the user.

In Okapi, there are four main sources of data used in query term weighting: collection frequency, term frequency, document length, and relevance feedback information [18]. The term weighting function BM25 used in Okapi was discussed in Section 2.2.3.

Okapi was set up specifically to provide an environment in which ideas could be tested by involving real users. Robertson, Walker and Beaulieu [21] wrote that the Okapi system has been used in a series of experiments on the TREC collections, investigating probabilistic models, relevance feedback, query expansion, and interaction issues. Although some of the Okapi projects involve laboratory type experiments, as exemplified by the involvement in the TREC, the main focus of Okapi-based projects has been to explore the inherent interactive nature of the retrieval process [20].

2.5 Query Operations

To improve the initial query formulation, which is treated as an initial attempt to retrieve relevant information, a variety of approaches could be deducted effectively. Query expansion (QE) is the process of reformulating a seed query to improve retrieval performance in IR operations [23]. The idea of relevance feedback is to involve the user into the retrieval process so as to improve the final result set. In particular, the user gives feedback on the relevance of documents in an initial set of results, which can go through one or more iterations [22].

In relevance feedback, users give additional input on documents (by marking documents in the results set as relevant or not), and this input is used to re-weight the terms in the query for documents. In query expansion, users give additional input on query words or phrases, possibly suggesting additional query terms.

2.5.1 Relevance Feedback and Pseudo-Relevance Feedback

Relevance feedback, as a process which uses judged relevance information to further improve IR performance, is one of the most effective query reformulation strategies [7]. It is believed that, for a specific query in a relevance feedback cycle, documents in the collection could be divided into the relevant set and the non-relevant set. The expected effect of relevance feedback is that the new query will be moved towards the relevant documents and away from non-relevant ones.

The relevance feedback process exploits the idea that it may be difficult for IR system users to formulate a good query when users don’t know the collection well, but it is easy to judge particular documents, and so it makes sense to engage in iterative query refinement [24]. In such a scenario, relevance feedback can also be effective in tracking the user’s evolving information need: seeing some documents may lead users to refine their understanding of the information they are seeking.

Pseudo-relevance feedback, also known as blind relevance feedback, provides a method for automatic local analysis [26]. It automates the manual part of relevance feedback, so that the user may get improved retrieval performance without an extended interaction. The method is to do normal retrieval to find an initial set of most relevant documents, to then assume that the top k ranked documents are relevant, and finally to do relevance feedback as before under this assumption [25].

This automatic technique mostly works [27]. It has been found to improve performance in the TREC ad hoc task, but it is not without the dangers of an automatic process. For example, if the query is about copper mines and the top several documents are all about mines in Chile, then there may be query drift in the direction of documents on Chile.

2.5.2 Query Expansion

Query expansion is the process of supplementing the original query with additional terms, and it can be considered as a method for improving retrieval performance. The initial query provided by the user may be an inadequate or incomplete representation of the user's information need, either in itself or in relation to the representation of ideas in documents [7]. Query expansion may take place in the initial query formulation or in the query reformulation stage, or in both.

Query expansion can be performed manually, automatically or interactively [27]. Automatic Query Expansion (AQE) occurs when the system selects appropriate terms for use in query expansion and automatically adds these terms to users' queries. In most cases, AQE is favored because the system can make a better selection of query expansion terms which have high probability of being relevant [27].

Conversely, Interactive Query Expansion (IQE) gives the control to the user who can mark documents and suggest potential query terms [27]. The benefit of IQE has been demonstrated that it can assist users arrive at an ideal query. Both AQE and IQE have been proved to be effective to get better results, but AQE is mostly based on collection independent knowledge structures and IQE mainly on dependent knowledge structures.

Chapter 3 Query Templates

To help IR system users articulate their complex information needs, we compared two query formulation interfaces in an IR system, and a novel approach to query formulation is proposed. To conduct this new approach, firstly we analyzed structure of TREC documents and topics guided by Rhetorical Structure Theory (RST), and then generated a query template in the form of fill-in-the-blanks questions.

3.1 Rhetorical Structure Theory

Rhetorical Structure Theory (RST) was originally developed as part of studies of computer-based text generation. RST offers an explanation of the coherence of texts that has led to areas of application beyond discourse analysis and text generation [27].The central constructs in RST are rhetorical relations. Text coherence is attributed principally to the presence of these relations which in RST are sufficient to analyse the majority of English texts and exceptions are only made for very unusual texts like poems and legal documents [28].

In ordinary usage, a text has a kind of unity that arbitrary collections of sentences lack. “RST, which is a framework for the description of texts in terms of functions and roles played by different segments of text, offers an explanation of the coherence of texts. For every part of a coherent text, there is some function, some plausible reason for its presence, evident to readers [29].” Consider the following sentences: “They're having a party next door. I couldn't find a parking space”. In this example, the second sentence represents evidence that the first sentence is true. Compare another example: “They're having a party next door. I like apples”. These two sentences lack coherence because it is not clear to the reader why they are placed together.

From the examples, we could observe that RST is intended to describe texts, rather than the processes of creating or reading and understanding them. “It posits a range of possibilities of structure -- various sorts of "building blocks" which can be observed to occur in texts [29].” The most frequent structural pattern is that two spans of text (virtually always adjacent, but exceptions can be found) are related such that one of them has a specific role relative to the other. A common case is a claim followed by evidence for the claim. RST posits an "Evidence" relation between the two spans. It also says that the claim is more essential to the text than the particular evidence, and this essentiality is represented by calling the claim span a nucleus and the evidence span a satellite. The order of spans is not constrained, but there are more likely and less likely orders for all of the relations. In the example, the relationship between two sentences is “evidence”. Satellite presents evidence that nucleus is true; reader believes satellite and writer wants reader to believe nucleus. If a relation does not have a particular span of text which is more central to the author’s purposes, it is called Multinuclear. An example is the neutral Contrast relation, whereby two nuclei are compared and contrasted.

As we discussed above, the relations are classified into two main types: nucleus-satellite and multinuclear. Nucleus-satellite relation can also be classified according to whether they are Presentational or Subject Matter [30]. “Presentational relations are those whose intended effect is to increase some inclination in the reader, such as the desire to act or the degree of positive regard for, belief in, or acceptance of the nucleus. Subject matter relations are those whose intended effect is to help the reader recognize the relations in question [29]. The presentational relations include: Antithesis, Background, Concession, Enablement, Evidence, Justify, Motivation, and Preparation; subject matter relations have Circumstance, Condition, Elaboration, Evaluation, Interpretation, Means, Non-volitional Cause, Non-volitional Result, Otherwise, Purpose, Solutionhood, Unconditional, Unless, Volitional Cause, Volitional Result [29]. In our study, we focus on identifying subject-matter relations.

RST is designed to enable the analysis of texts. The analyst can make claims based on the definition of the relations and other structures of RST, and the analysis consists of explicit claims by accounting for the sense of unity, connectedness and coherence of ordinary written monologues. “It is the possibility of enumerating the specific claims of the analysis which makes RST analyses comparable to other approaches [28].”

3.2 TREC Documents and Topics

TREC collection has been growing steadily over the years. The documents come from a wide range of sources, including Wall Street Journal, Financial Times, US Patent, San Jose Mercury News, and so on. The major structures such as a field of the document number and a field of the document text are common to all documents in TREC collections. Minor structures might be different across collections since NIST decided to preserve as much of the original structures as possible while providing a common framework [31].

An example of a TREC document is shown as below:

<DOC>

<DOCNO>WSJ880406-0090</DOCNO>

<HL> AT&T Unveils Services to Upgrade Phone Networks Under Global Plan </HL>

<AUTHOR> Janet Guyon (WSJ Staff) </AUTHOR>

<DATELINE> New York </DATELINE>

<TEXT> American Telephone & Telegraph Co. introduced the first of a new generation of phone services with broad implications for computer and

communications equipment markets….

</TEXT>

</DOC>

The TREC collection includes a set of example information requests which can be used for testing a new ranking algorithm, which is referred to as a topic as we discussed above. An example of an information request, a description of an information need in natural language, is illustrated below:

<top>

<num> Number: 303

<title> Hubble Telescope Achievements

<desc> Description:

Identify positive accomplishments of the Hubble telescope since it was launched in 1991.

<narr> Narrative:

Documents are relevant that show the Hubble telescope has produced new data, better quality data than previously available, data that has increased human knowledge of the universe, or data that has led to disproving previously existing theories or hypotheses.

Documents limited to the shortcomings of the telescope would be irrelevant.

Details of repairs or modifications to the telescope without reference to positive achievements would not be relevant.

</top>

A number of the topics are prepared for each TREC conference. The topics are written by people who were experienced users of real systems and represent their information needs [31].

Vechtomova has manually analyzed the descriptions and narratives of 100 TREC topics (#301-400) from TREC 6 and 7 ad-hoc tracks using the Rhetorical Structure Theory [40]. Based on this analysis, a set of domain- and topic-independent semantic relations between concepts to fit in users' information needs was identified in this study.

3.3 Question Templates

Although in some situation users may be unable to articulate their information needs because they lack the knowledge or the vocabulary to describe them, users still should be considered as the useful source of terms for query expansion because the traditional interfaces for querying and relevance feedback are not optimal for eliciting the robust and useful description [2].

3.3.1 Related Work

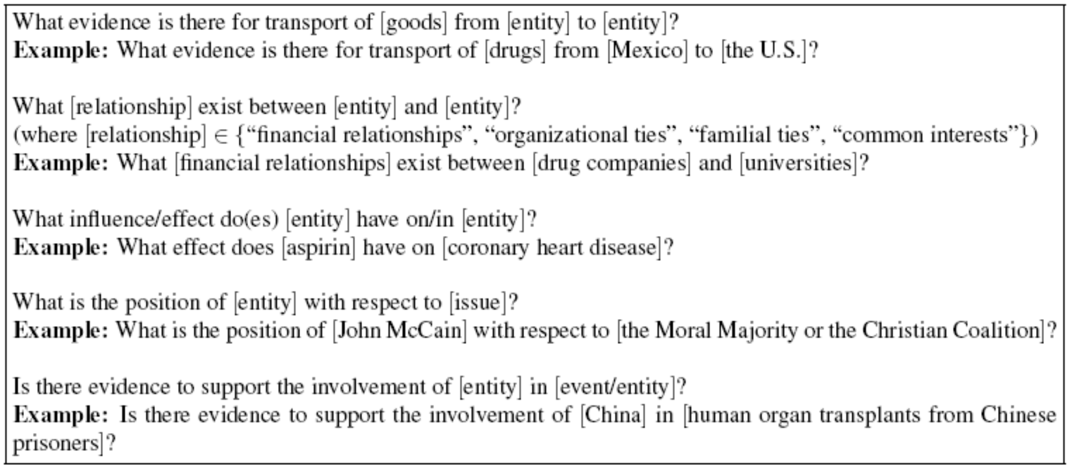
In TREC 2006, the complex information needs were explored by ciQA Track, which represented an extension and refinement of so-called "relationship" questions [3]. The concept of a "relationship" is defined as the ability of one entity to influence another, including both the means to influence and the motivation for doing so.

In the ciQA task, a relationship question is composed of two parts: Template and Narrative [3]. Specifically, the question template is a stylized information need that has a fixed structure and free slots (items in square brackets) whose instantiation varies across different topics. The narrative is a free-form natural language text that elaborates on the information need [3]. For example:

> **Template:** What evidence is there for transport of [drugs] from [Mexico] to [the U.S.]?
>
> **Narrative:** The analyst would like to know of efforts to curtail the transport of drugs from Mexico to the U.S.

The five template types were developed for the ciQA task shown in Figure 3.1.

Figure 3.1 The five templates used in the TREC 2006 ciQA task



What evidence is there for transport of [goods] from [entity] to [entity]?
Example: What evidence is there for transport of [drugs] from [Mexico] to [the U.S.]?

What [relationship] exist between [entity] and [entity]?
(where [relationship] $\in$ {“financial relationships”, “organizational ties”, “familial ties”, “common interests”})
Example: What [financial relationships] exist between [drug companies] and [universities]?

What influence/effect do(es) [entity] have on/in [entity]?
Example: What effect does [aspirin] have on [coronary heart disease]?

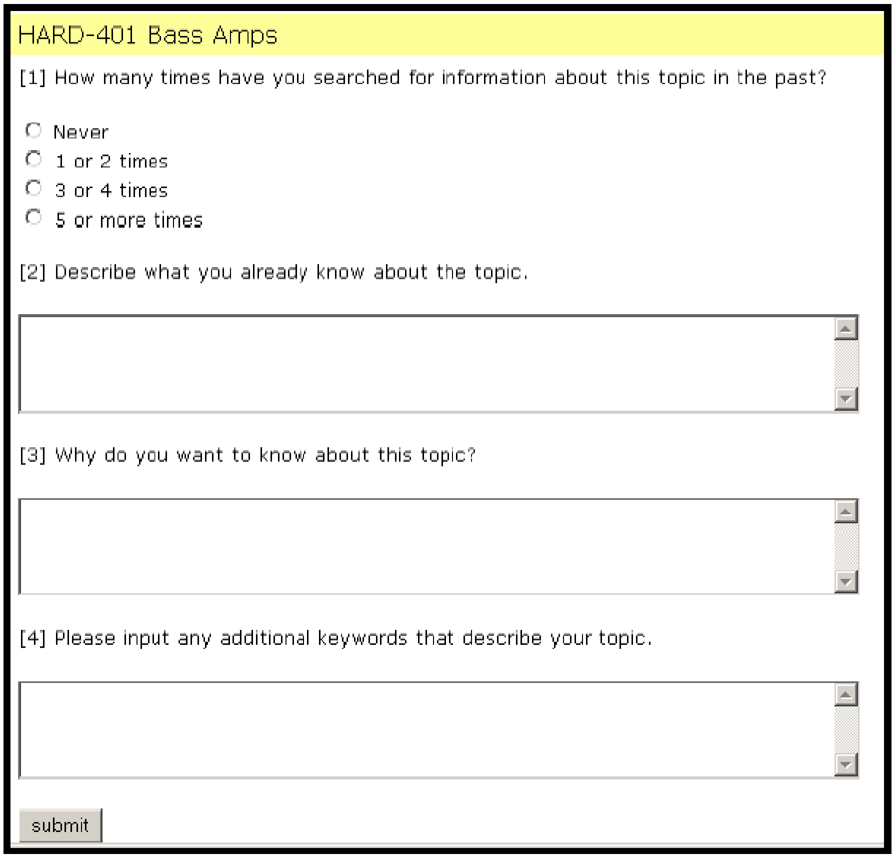
What is the position of [entity] with respect to [issue]?
Example: What is the position of [John McCain] with respect to [the Moral Majority or the Christian Coalition]?

Is there evidence to support the involvement of [entity] in [event/entity]?
Example: Is there evidence to support the involvement of [China] in [human organ transplants from Chinese prisoners]?

The TREC HARD (High Accuracy Retrieval from Documents) track protocol included a one-time interaction with the users (NIST assessors), whereby upon receiving an initial query statement, participating sites could send to the user a clarification form, which could ask for any information about the user's information need. Kelly et al. [2] conducted an experiment within the framework of the HARD track of TREC 2004. In this experiment, a template form, which helps users to supply more information about their information needs, was implemented. Similar to the HARD track’s clarification forms, which allowed participants to elicit information from users through a single interaction, this template form interaction consisted of users completing forms [3].

Kelly et al. [2] presented to users a generic and document-independent feedback form after initial querying in the hope of eliciting more complete descriptions of users' information needs. The information obtained from the form was treated as a source of terms for query expansion, a series of experimental runs based on this information were conducted. This particular clarification form that consisted of four questions and could be used for all topics without modification is displayed below in Figure 3.2.

Figure 3.2 Clarification Form [2]



HARD-401 Bass Amps

[1] How many times have you searched for information about this topic in the past?

○ Never
○ 1 or 2 times
○ 3 or 4 times
○ 5 or more times

[2] Describe what you already know about the topic.

[3] Why do you want to know about this topic?

[4] Please input any additional keywords that describe your topic.

The first question was designed to assess the topic familiarity by asking users to indicate how many times they had searched for information about their topics in the past. Questions 2, 3, 4 were designed to elicit information from users about their topics for query expansion. As shown above, these questions used large text boxes that allowed users to view the entirety of their responses and type in longer responses than they would if presented with a short line. Questions 2 and 3 were open-ended questions and encouraged users to respond with natural language. Question 4 asked users to list any additional keywords describing their topics in order to help users to make a selection of good terms.

All the results of the experiment provided strong evidence for the effectiveness of runs comprised of a combination of representations of the users' information needs [2]. These results suggested that when the information elicited from each question was used in combination for query expansion, it was most useful. Indeed, probing users with a number of different but related questions from the clarification form might elicit the most robust and useful problem descriptions [2]. On the other hand, the overall performance results also suggested a strong relationship between query length and performance for the experimental techniques.

3.3.2 Constructing Query Templates

To explore the query templates, Vechtomova manually analyzed the descriptions and narratives of 100 TREC topics (#301-400) from TREC 6 and TREC 7 ad hoc tasks [40] by using RST. By reading the text of the description and narrative for a topic, Vechtomova dropped five infrequent relations: “Circumstance”, “Condition”, “Unconditional”, “Unless”, and “Contrast”, which is a kind of multinuclear relations, and for the other relations remaining, the number of topic instances of each relation was recorded [40]. The analysis result is shown in Table 3.1 below. As the final output, a set of domain and topic independent semantic relations between entities was created.

Table 3.1 Categories and Topics (301-400)

Category	Topics
1. Cases, instances of something	301, 302, 304, 307, 309, 311, 313, 317, 322, 323, 325, 326, 327, 328, 332, 339, 342, 343, 344, 346, 348, 353, 354, 357, 358, 361, 362, 363, 364, 365, 366, 367, 374, 376, 378, 383, 389, 393, 394, 395, 397
2. Statistical data on something	306, 315, 321, 336, 358, 398
3. Cause	333, 336, 369, 396, 397
4. Effect	310, 309, 337, 338, 345, 350, 352, 358, 365, 372, 390, 391, 396, 398
5. Use (application) of something	313, 314, 355, 366, 382, 388, 392
6.1 Measures & solutions to combat/alleviate/reduce something	329, 335, 336, 337, 340, 341, 344, 347, 350, 364, 369, 370, 380, 381, 383, 398
6.2 Measures & solutions to to improve/increase/facilitate something	
7. Opinion	309, 316, 334, 379
8. Positive & Negative aspects	308, 356, 357, 360, 371, 372, 379, 380, 385
8.1. Only positive aspects	303
8.2. Only negative aspects	331, 335
9. Comparison and contrast	305, 318, 339, 398
10. In-depth information about something	309, 312, 316, 319, 320, 321, 330, 336, 337, 351, 355, 359, 368, 372, 373, 375, 377, 384, 385, 390, 393, 396, 398
11. Impediment	384
12. Goal, purpose	384
13. Method/procedure	386, 387

To represent the templates with the fill-in-the-blanks question format, some key phrases from the TREC topic authors' actual questions were collected. Based on the different relations, one to two questions were generated for each relation. In total, 13 query templates were written representing general semantic relations between concepts. The templates are listed as below:

1. Cases, instances of something [Find all cases/instances of X.]
2. Statistical data on something [How often does X occur? How many X happened?]
3. Cause [Find what causes X. Find about why/how Y causes X.]
4. Effect [Find what effect X has on Y]

5. Use (application) of something [Find existing/potential uses/applications of X. Find out about the application of X to Y.]
6. Measures & solutions

    6.1. to combat/alleviate/reduce something [Find what has been done (or could be done) to combat/alleviate/reduce X (or the effect of X).]

    6.2. to improve/increase/facilitate something [Find what has been done (or could be done) to improve/increase/facilitate X).
7. Opinion [Find what people think about X.]
8. Positive & negative aspects [Find positive & negative aspects/evidence of X.]

    8.1. Only positive aspects [Find positive aspects of X.]

    8.2. Only negative aspects [Find negative aspects (problems) of X.]
9. Comparison and contrast [Compare X and Y. Find which is the best/worst X.]
10. In-depth information about something [Find everything about X.]
11. Impediment [Find what impedes (gets in the way of) X]
12. Goal, purpose [Find what is the goal/purpose of X.]
13. Method/procedure [Find what methods/procedures are (can be) used to do/achieve X.]

As can be seen from the above discussion, there is no one-to-one mapping between the RST relations and the developed templates [32]. For example, “Non-volitional Cause” [Nucleus (N) is caused by Satellite (S), where S is a non-volitional action] and “Volitional Cause” [N is caused by S, where S is a volitional action] relations are represented by one template “What causes ____?” Moreover, two templates represent the relation “Otherwise” [realization of N prevents realization of S]: “What has been done (could be done) to combat/alleviate the effect of ____?” and “What impedes (gets in the way of) ____?” The system users are expected to interact with these templates by filling in those they think relevant and helpful to achieve their information requirements.

Although the RST relations and fill-in-the-blanks questions do not represent all possible entity relations about which users may want to find information, this set of relations is still broad enough to investigate the research questions posed by our study [32].

Chapter 4 User Study

In the design of IR search systems, the importance of understanding the users' strategies has been acknowledged by many researchers. To control the complexity of our user-centered evaluation methodology, we designed two search user interfaces: control interface and experimental interface. Our system was designed with the interactive PHP/MySQL based interface to Okapi system developed by Susan Jones et al. [43]. The experiment was conducted using Latin Square principle with the 30 graduate students' participation.

4.1 Experimental Design

To reach the major goal of our user study, which is to investigate whether our IR system users can articulate complex information needs by filling in topic-independent templates, we explored two research questions:

> RQ1: Are users able to articulate complex information needs using topic-independent query template?
>
> RQ2: When used with a bag-of words retrieval system, do the queries formed using template-based interface lead to better search results than queries formed using a single-textbox interface?

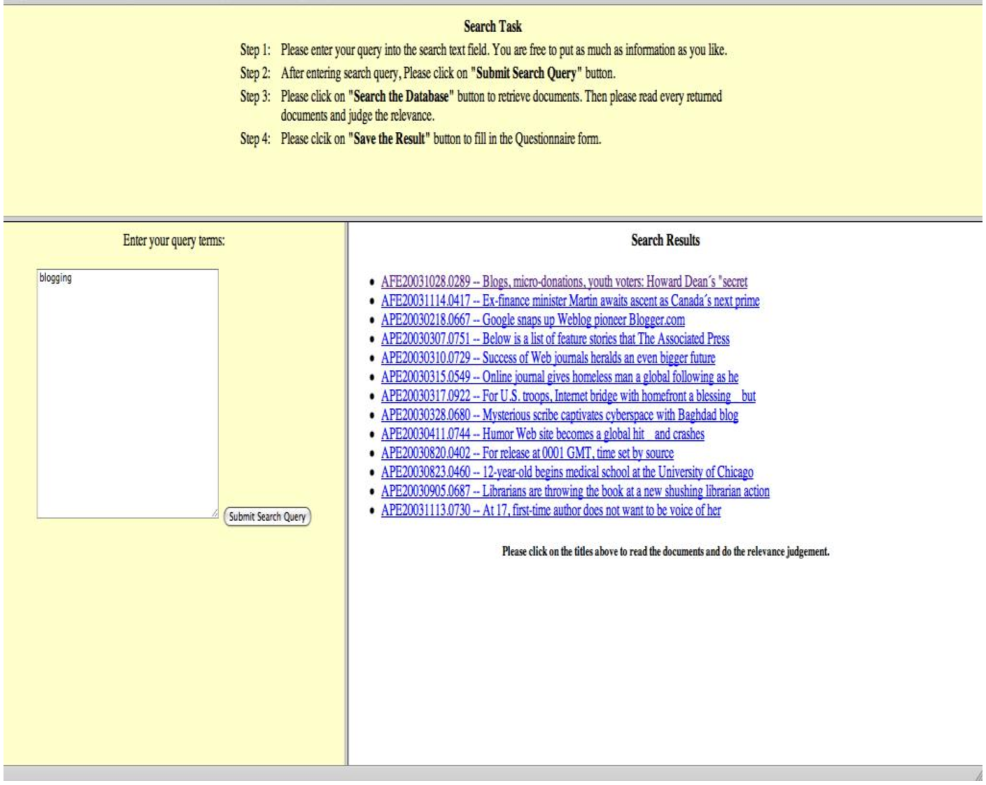
To investigate these research questions, we gathered and analyzed the following types of data gathered during our user study:

- IR Search session logs. We recorded user sessions related data, including time to read given documents, time to formulate the query, number of query terms entered, number of templates filled in, and time to judge the document relevance.
- Retrieval performance values based on the users' binary relevance judgments. Users' judgments about 15 retrieved documents in response to the users' queries were calculated with Precision at 5, 10, and 15.
- Users' subjective comments. During the user study, every user was asked to fill in a questionnaire, in which satisfaction with the search results, satisfaction with the overall system, familiarity with the topic prior to search, and user background were elicited.

We use the IR search session log data to measure the users' ability to formulate queries by using templates. The retrieval performance values and users' reported satisfaction levels were used to measure the retrieval effectiveness of template-based query formulation. Moreover, to analyze the users' interaction with IR system for some topics, we combine all these three types of data together.

To evaluate our proposed method to gather user searching queries, two search interfaces were developed: the control and the experimental (template-based). The control interface contained one large textbox, where the user was encouraged to enter more words, phrases and even sentences representing his/her information need (Figure 4.1). The experimental interface consisted of 13 topic-independent query templates (Figure 4.2). Each template follows the fill-in-the-blanks format. The IR system used with both interfaces in our user study was Okapi weighted function BM25 with the default parameter values b as 0.75 and k1 as 1.2 [33]. K1 and b are two constants in probabilistic relevance weight functions.

Figure 4.1 The control system interface



Search Task

Step 1: Please enter your query into the search text field. You are free to put as much as information as you like.

Step 2: After entering search query, Please click on "**Submit Search Query**" button.

Step 3: Please click on "**Search the Database**" button to retrieve documents. Then please read every returned documents and judge the relevance.

Step 4: Please clcik on "**Save the Result**" button to fill in the Questionnaire form.

Enter your query terms:

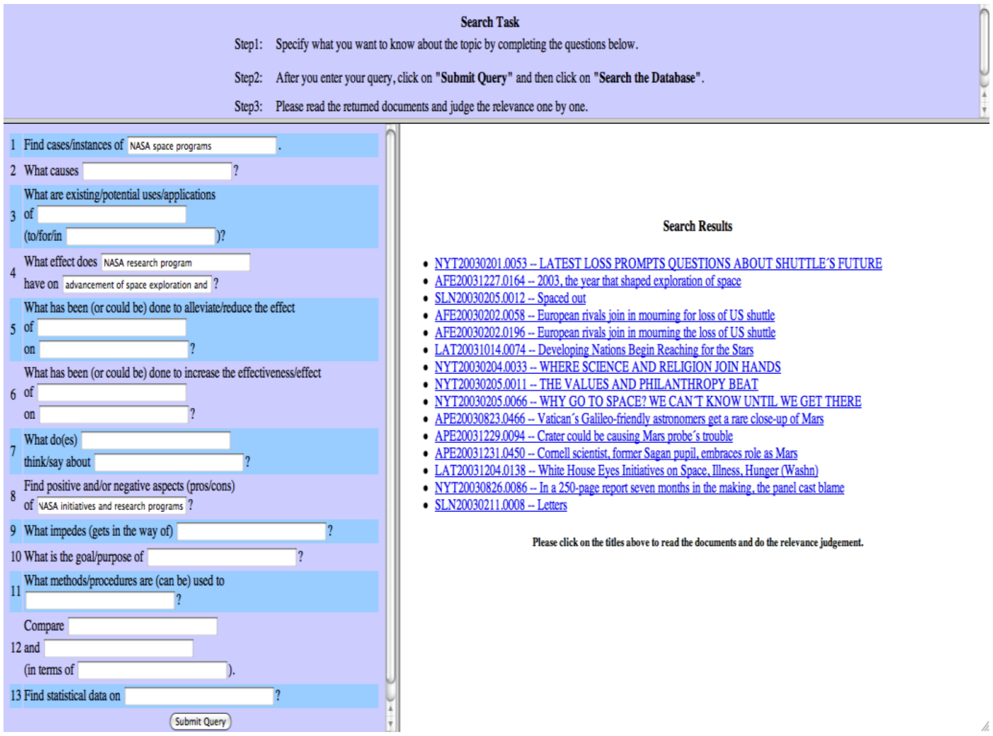
blogging

Search Results

- AFE20031028.0289 -- Blogs, micro-donations, youth voters: Howard Dean´s "secret
- AFE20031114.0417 -- Ex-finance minister Martin awaits ascent as Canada´s next prime
- APE20030218.0667 -- Google snaps up Weblog pioneer Blogger.com
- APE20030307.0751 -- Below is a list of feature stories that The Associated Press
- APE20030310.0729 -- Success of Web journals heralds an even bigger future
- APE20030315.0549 -- Online journal gives homeless man a global following as he
- APE20030317.0922 -- For U.S. troops, Internet bridge with homefront a blessing but
- APE20030328.0680 -- Mysterious scribe captivates cyberspace with Baghdad blog
- APE20030411.0744 -- Humor Web site becomes a global hit and crashes
- APE20030820.0402 -- For release at 0001 GMT, time set by source
- APE20030823.0460 -- 12-year-old begins medical school at the University of Chicago
- APE20030905.0687 -- Librarians are throwing the book at a new shushing librarian action
- APE20031113.0730 -- At 17, first-time author does not want to be voice of her

Please click on the titles above to read the documents and do the relevance judgement.

Figure 4.2 The experimental (template-based) system interface



Search Task

Step1: Specify what you want to know about the topic by completing the questions below.

Step2: After you enter your query, click on "**Submit Query**" and then click on "**Search the Database**".

Step3: Please read the returned documents and judge the relevance one by one.

1 Find cases/instances of NASA space programs .

2 What causes ?

3 What are existing/potential uses/applications of (to/for/in)?

4 What effect does NASA research program have on advancement of space exploration and ?

5 What has been (or could be) done to alleviate/reduce the effect of on ?

6 What has been (or could be) done to increase the effectiveness/effect of on ?

7 What do(es) think/say about ?

8 Find positive and/or negative aspects (pros/cons) of NASA initiatives and research programs ?

9 What impedes (gets in the way of) ?

10 What is the goal/purpose of ?

11 What methods/procedures are (can be) used to ?

12 Compare and (in terms of).

13 Find statistical data on ?

Search Results

- NYT20030201.0053 -- LATEST LOSS PROMPTS QUESTIONS ABOUT SHUTTLE´S FUTURE
- AFE20031227.0164 -- 2003, the year that shaped exploration of space
- SLN20030205.0012 -- Spaced out
- AFE20030202.0058 -- European rivals join in mourning for loss of US shuttle
- AFE20030202.0196 -- European rivals join in mourning the loss of US shuttle
- LAT20031014.0074 -- Developing Nations Begin Reaching for the Stars
- NYT20030204.0033 -- WHERE SCIENCE AND RELIGION JOIN HANDS
- NYT20030205.0011 -- THE VALUES AND PHILANTHROPY BEAT
- NYT20030205.0066 -- WHY GO TO SPACE? WE CAN´T KNOW UNTIL WE GET THERE
- APE20030823.0466 -- Vatican´s Galileo-friendly astronomers get a rare close-up of Mars
- APE20031229.0094 -- Crater could be causing Mars probe´s trouble
- APE20031231.0450 -- Cornell scientist, former Sagan pupil, embraces role as Mars
- LAT20031204.0138 -- White House Eyes Initiatives on Space, Illness, Hunger (Washn)
- NYT20030826.0086 -- In a 250-page report seven months in the making, the panel cast blame
- SLN20030211.0008 -- Letters

Please click on the titles above to read the documents and do the relevance judgement.

There were thirty students, all graduate students of University of Waterloo, participating in the study. Based on the short entry questionnaire (Appendix C), the user demographic details are summarized in Table 4.1 below.

Table 4.1 User Characteristics

Question	Number of responses
Age group	
19 and under	0
20 – 29	25
30 – 39	4
40 – 49	1
50 – 59	0
Gender	
Male	21
Female	9
How often do you use search engines?	
Every day	27
Several times per week	3
Less than once a week	0
Very rarely	0
What do you use search engines for?	
Work/Study	3
Personal	1
Both	26
Do you use advanced search engine functions?	
Yes	16
No	14

Our search topics were selected from the TREC HARD track 2004 collection, which consists of 635,650 documents [34]. As the search tasks, we used 10 topics, which were not the same as those used to create the templates using RST analysis. The topics represent complex information needs, which are considered as cases where the topic narrative contains at least one general semantic relation. Moreover, in order to eliminate topics with insufficient coverage in the collection, we selected the topics having 50 or more relevant documents based on the TREC2004 result sets. All the topics and the number of relevant documents in the HARD track 2004 are given in table 4.2.

Table 4.2 Ten selected search topics from TREC HARD track 2004

	Topic ID	Number of related documents	Topic Title
1	406	75	The Diamond Industry
2	409	318	AIDS in Africa
3	418	178	Immigration Post 9-11
4	421	168	Software Patents
5	422	137	Video game crash
6	425	104	Cancun World Trade Talks
7	432	111	Farm Subsidies
8	441	521	European Union
9	442	284	Interest Rates
10	443	204	Hand-Held Electronics

An example of the selected topics is shown in table 4.3.

Table 4.3 Example of the selected TREC topic

Topic Number: 406
Title: The Diamond Industry
Narrative: What levels of corruption exist in the Diamond Industry and how does the development of a synthetic flawless diamond impact the industry? The diamond industry has long been associated with corruption on many levels. With the emergence of a new synthetic stone, the industry finds itself at a crossroads. Virtually anything dealing with corruption in the diamond industry and the effects generating from the development of the synthetic gem is on topic. Articles relating to the technical aspects of cutting and setting diamonds are off topic. Scientific articles are also off topic.

Each participant used both user interfaces, which required us to prepare two balanced task sets. The use of task sets was counterbalanced between the user interfaces and participants using Latin Square. Each user performed two search tasks, one with the control interface and one of the topics, and the other with the experimental interface and a different topic. The order of the systems was rotated. For example, two topics A and B were assigned to six users as shown in Table 4.4. In order to control the effect of topic on the retrieval performance, we assigned the same topic to two users with the control system and experimental system equally. We used the same method to assign the search tasks to all 30 users.

Table 4.4 Experiment design

User ID	Search Task 1		Search Task 2	
	Topic	Interface	Topic	Interface
1	A	Control	B	Experimental
2	A	Experimental	B	Control
3	B	Control	A	Experimental
4	B	Experimental	A	Control
5	A	Control	B	Experimental
6	A	Experimental	B	Control

4.2 Experimental Protocol

All participants were recruited by sending recruiting e-mail (Appendix A) with the attachment of user study information letter (Appendix B). Once participants confirmed the participation and reserved the time slot, the IR system was set up and tested to be ready for conducting the user study. The experiment received full ethics clearance from the Office of Research Ethics at University of Waterloo (Appendix E).

Before the user started their search tasks, each user read and signed the informed consent letter (Appendix B) and was given a short entry questionnaire, followed by a brief tutorial of the first system (experimental or control). To avoid the bias and simulate the real search tasks, the tutorial systems were designed to be the same as the experimental and control systems respectively with the different search topics, also from

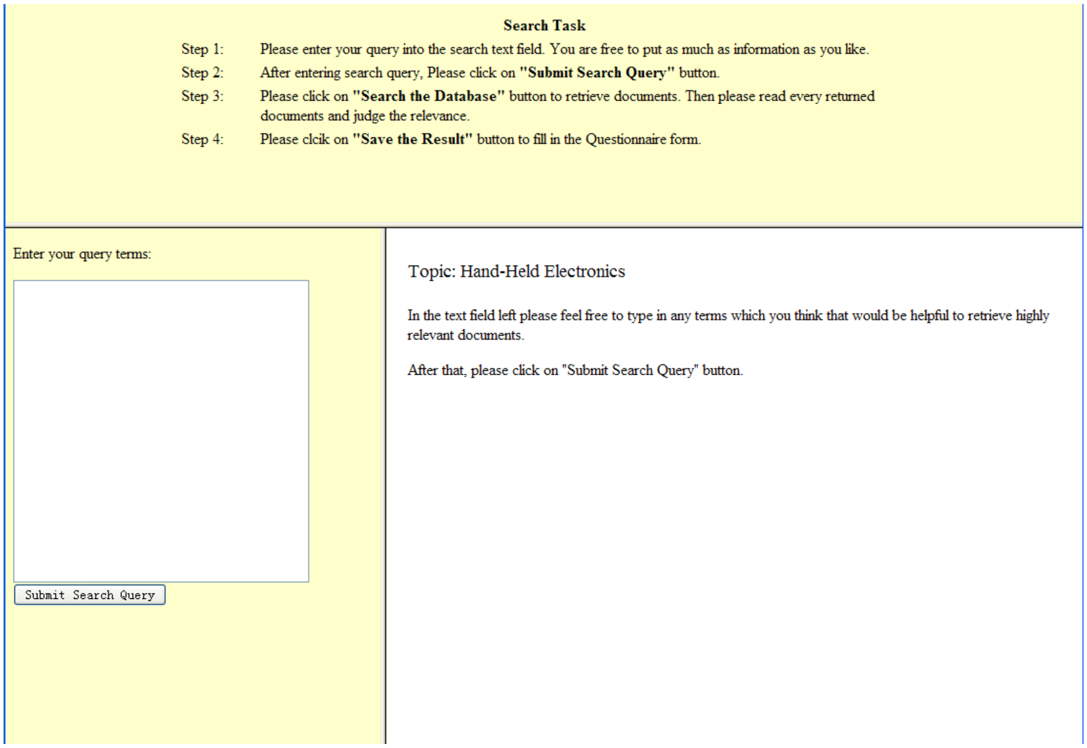
TREC HARD collection. The same script and procedures were used during the tutorial and users were allowed to ask questions when the tutorial finished.

The participants were then shown the TREC topic title, description, and narrative of their assigned topics on paper, on which they could make notes. Once the participant was ready to start, the experiment system was launched on the screen, and he/she could logon to the system by using assigned user ID. Then, he/she was presented with their first search task's topic information which was exactly same as the information they had on the paper. After the reading of the title, description, and narrative of the topic, he/she was presented with 3 relevant documents for this topic, which were randomly chosen from HARD2004 relevance judgments files. The reason to show these 3 documents is to increase their level of familiarity with the topic based on the hypothesis that template-based interface would be helpful to expert users with complex information needs. Such users should already have a high level of familiarity with the subject and they are eager to learn about specific relationships between concepts or entities in their domain of interest. On the contrary, those users with little or no familiarity of the topic are unlikely to benefit from the use of query templates, as he/she may simply have insufficient knowledge to identify related concepts and thus formulate complex queries. By giving these three documents, we helped the participants to increase their familiarity with their search topics, so that they may be interested in getting some specific information during their search tasks.

After the user finished reading these 3 documents, he/she was then presented with the corresponding query formulation interface (experimental or control) and was asked to formulate a query based on the given topic. As we can see from Figure 4.3 and 4.4, the text of topic was not displayed on the screen at the query formulation stage to prevent copying and pasting directly. The user still can use the printed paper to review the topic.

After the user submitted the query, the query terms were stemmed, and document retrieval was performed using Okapi BM25. We returned the top 15 retrieved documents on the screen by excluding duplicate documents and the three initially given documents from the retrieved set. The user then judged each of them as “relevant” or “non-relevant”. At the end of each search session, he/she was asked to fill in a questionnaire eliciting feedback (Appendix F, Appendix G) regarding the system used for that task. After a short break, the user proceeded to the second task following the same experimental principles as in the first. Some discussions with the user about his/her search task, system design and IR topic concluded the experiment session.

Figure 4.3 Control interface query formulation



Search Task

Step 1: Please enter your query into the search text field. You are free to put as much as information as you like.

Step 2: After entering search query, Please click on **"Submit Search Query"** button.

Step 3: Please click on **"Search the Database"** button to retrieve documents. Then please read every returned documents and judge the relevance.

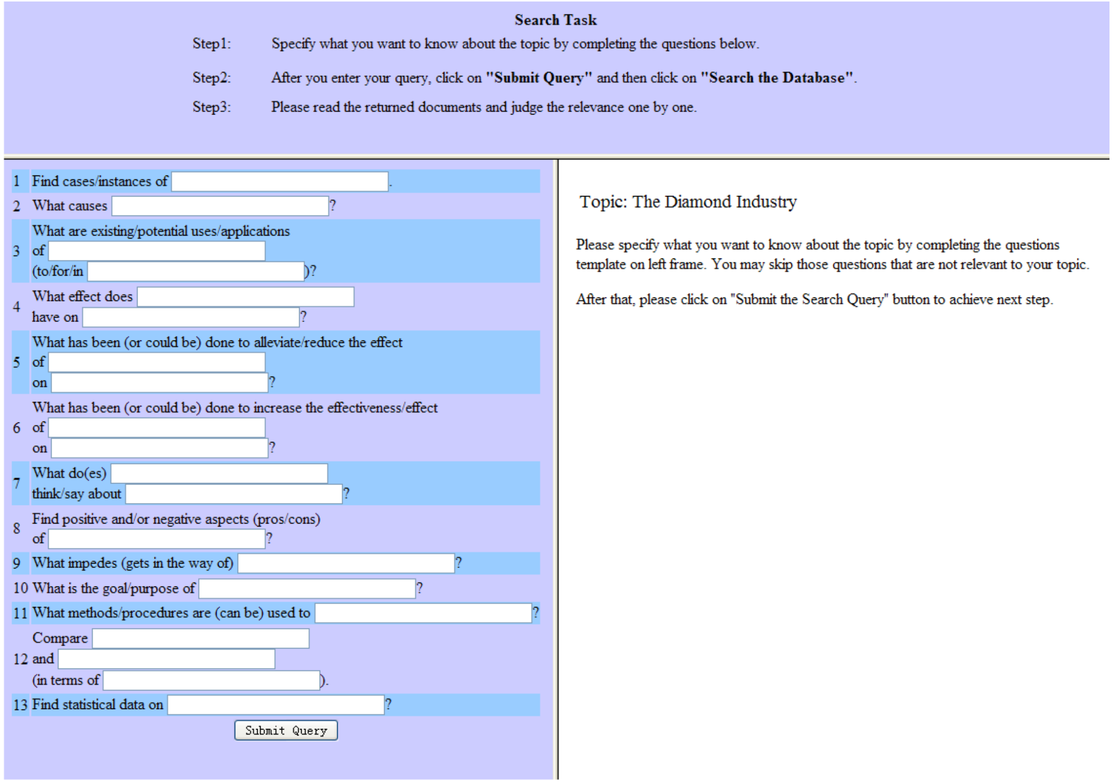
Step 4: Please clcik on **"Save the Result"** button to fill in the Questionnaire form.

Topic: Hand-Held Electronics

In the text field left please feel free to type in any terms which you think that would be helpful to retrieve highly relevant documents.

After that, please click on "Submit Search Query" button.

Figure 4.4 Experimental interface query formulation



Search Task

Step1: Specify what you want to know about the topic by completing the questions below.

Step2: After you enter your query, click on **"Submit Query"** and then click on **"Search the Database"**.

Step3: Please read the returned documents and judge the relevance one by one.

1 Find cases/instances of ______.

2 What causes ______?

3 What are existing/potential uses/applications of ______ (to/for/in ______)?

4 What effect does ______ have on ______?

5 What has been (or could be) done to alleviate/reduce the effect of ______ on ______?

6 What has been (or could be) done to increase the effectiveness/effect of ______ on ______?

7 What do(es) ______ think/say about ______?

8 Find positive and/or negative aspects (pros/cons) of ______?

9 What impedes (gets in the way of) ______?

10 What is the goal/purpose of ______?

11 What methods/procedures are (can be) used to ______?

12 Compare ______ and ______ (in terms of ______).

13 Find statistical data on ______?

Topic: The Diamond Industry

Please specify what you want to know about the topic by completing the questions template on left frame. You may skip those questions that are not relevant to your topic.

After that, please click on "Submit the Search Query" button to achieve next step.

Chapter 5 Results and Analysis

Two sets of experiments were designed to meet the goal of our study. The results help us to study template-filling as a query formulation method. The analysis of these results is discussed below.

5.1 Retrieved Performance

Based on user relevance judgments results, we evaluated the performance of two interfaces which were designed to articulate users' complex information needs by formulating search queries. The performance of the template-based and the control interfaces was measured using precision at 5, 10 and 15 documents. The results are summarized in Table 5.1.

Table 5.1 Performance of the queries formed using the control and template-based interfaces (* is statistically significant; t-test, p=0.037)

Interface	**P@5**	**P@10**	**P@15**
Control	0.5133	0.4767	0.4667
Template-based	0.6467	0.6067*	0.5533

As shown in the above table, the template based interface always led to higher performance than control interface. The t-test result also showed that the improvement of performance between control and experimental systems is statistically significant. The performance of the queries formed using the template based interface measured in P@5 is 25.97% higher than the control, in P@10 it is 27.27% higher and in P@15 it is 18.57% higher.

5.2 Query and search session characteristics

In our user study, we have logged different parameters during all the user search sessions, namely: time spent on reading the 3 sample relevant documents shown to users prior to each search session, time spent on formulating the query, time spent on reading and judging the relevance of the top 15 retrieved documents, number of query terms entered (tokens), number of unique query terms entered (types) excluding stopwords, and inverse document frequency (idf) weight of the query terms entered. The average values of these parameters are summarized in table 5.2.

Table 5.2 Characteristics of the search sessions using control and template-based interface

Logged Experiment Parameters	**Control (mean±stdev)**	**Template-based (mean±stdev)**
Time spent on reading 3 sample relevant documents (min:sec)	09:28±06:10	08:18±05:31
Time spent on formulating the query (min:sec)	04:08±03:39	04:35±03:36
Time spent on reading & judging 15 retrieved documents (min:sec)	22:28±14:22	22:09±12:28
Number of all query terms (tokens) including stopwords	16.17±9.9	31.83±19.36
Number of unique query terms (types) excluding stopwords	10.2±5.68	12.03±5.15
idf of non-stopword query terms	4.05±1.09	3.96±0.78

As we can see, the average time users spent on reading 3 sample relevant documents in control interface session is about one minute longer than it in template-based interface. With the consideration of many subjective factors including the participants' language background, reading skills, and the familiarity with the given search topics, this difference is not significant. Moreover, based on the statistical records to all given sample documents, we know the average words including the title of each document is about 400. Since our users are all graduate students from University of Waterloo, this actually proved that our users read the documents very carefully and they spent twice as much time, as they read with their normal reading speed based on the reading skills definition at http://www.readingsoft.com/ [35], which states that as a good reader, he/she could read 300 words per minute on screen. This evidence of careful reading is useful to suggest that users increased their familiarity with the search topics.

As we can observe from the log of the time spent on formulating the query, the users spent on average 27 seconds more when using the templates, compared to entering the query into control interface. This observation that users spent more than 4 minutes on both interfaces, can provide further support to the discussion Kelly, Dollu, and Fu had in 2005 that users were willing to provide lengthy responses to articulating their information needs precisely, if the system encourages them to do so [2]. The longer time spent on the experimental system is not surprising since we expected that the templates require longer time to read and think whether and how each fill-in-the-blanks question relates to the user's information need.

The observation of the time spent reading and judging 15 retrieved documents for both interfaces are almost the same, as we expected. The average time spent on reading one document is 90 and 88 seconds for the control and experimental system respectively. Based on the definitions of the reading skills [35] and our participants' demographic category – graduate students, this result shows that the users read the documents with their usual reading speed.

Another interesting observation from the above table is the number of query terms entered. The users on average entered 16.17 query term tokens (including stopwords) when using the control interface. This number increased considerably to 31.83 when using the template-based interface. The difference (96.91%) is statistically significant (t-test, p=0.0002). Without duplicate words and stopwords, the users on average entered 10.2 query term tokens in control system and 12.03 tokens in template-based system.

The difference (17.97%) is not statistically significant. Interestingly, on average 19.8 query terms, which is 62.2% of all tokens entered using the experimental interface, are either duplicates of previously entered words for the same query by the user, or stopwords. At the same time, in the control interface, there are only 5.97 such words on average which is 36.92% of all tokens. This is not entirely unexpected since the user may be interested in different aspects of the same concept and entered it into several templates. Also, when filling in the templates, it is very possible that users had to put some words, that are considered stopwords (e.g. “or”, “in”, “by”, “and”, “of”), in order to make the sentences coherent and grammatically correct. Nevertheless, from the table we still can easily discover that the users entered on average more unique non-stopword query terms using the template-based interface, which may be one of the reasons for its higher retrieval performance compared to the control interface. On the other hand, the average idf of query terms is slightly higher (4.4%) for the control interface, but the difference is not statistically significant.

Table 5.3 Templates and RST relations

RST relation	Template	$N_{completed}$
Elaboration	Find cases/instances of _____.	28
Elaboration	Find statistical data on _____?	16
Cause (volitional/ non-volitional)	What causes ____?	21
Result (volitional/ non-volitional)	What effect does _____ have on _____?	17
Purpose	What are existing/potential uses/applications of _____ (to/for/in____)?	23
Purpose	What is the goal/purpose of _____?	15
Otherwise	What impedes (gets in the way of) _____?	17
Otherwise	What has been done (could be done) to alleviate/reduce the effect of _____ on _____?	20
Means	What has been done (could be done) to increase the effectiveness/effect of _____ on _____?	18
Evaluation	What do(es) _____ think/ say about _____?	17
Evaluation	Find positive and/or negative aspects (pros/cons) of _____?	18
Solutionhood	What methods/procedures are (can be) used to _____?	15
Contrast	Compare _____and _____ (in terms of _____).	7

In Table 5.3 we list the question templates and their corresponding RST relations. The $N_{completed}$ is the number of users who completed each template in our user study. The average number of templates filled in by the users is 7.8 (standard deviation 3.13), which suggests that users were able to interpret and understand the templates in the context of their assigned search task. The RST relations corresponding to the most frequently filled in templates are “Elaboration”, “Purpose”, and “Otherwise”. The relation corresponding to the least frequently filled in template is “Contrast”.

As we discussed the experimental protocol above, in our experiment every TREC topic was assigned to three participants using template-based system. To assess the reliability of agreement among the users with respect to the templates they filled in for the assigned topic, we calculated the Fleiss Kappa as 0.14, which corresponds to slight agreement. The reason why no strong agreement was observed is probably due to some flexibility in the assigned search tasks: while the users were given rather detailed topic descriptions, they were able to interpret them somewhat differently based on their previous knowledge of the subject, focusing on different aspects of the topic, and therefore, filling in different templates. In chapter 6, we discuss some cases about how users filled in the templates and interacted with template-based system.

5.3 User satisfaction

During our user study, after each search session, each user was asked to complete a short exit questionnaire designed to elicit subjective feedback regarding the search session and the interface. The questionnaires designed for two search sessions were exactly same with 7 questions (as given in Appendix F for control interface and Appendix G for experimental interface). There was one open-ended question “Please describe difficulties (if any) in formulating the query” and six close-ended questions, each with four answers to choose from, for example:

Q6. How satisfied are you with the search results?

- Very satisfied. I found most of the information I was searching for.
- Somewhat satisfied. I found some relevant information.
- Not too satisfied. Most of the sought information was not found.
- Not satisfied at all. None of the sought information was found.

To quantitatively analyze the users' feedback, we mapped the responses to the questions to a 4-point scale with 1 being the most negative response and 4 – the most positive. The average responses are given in Table 5.4 below.

Table 5.4 User questionnaire analysis results

	Control (mean±stdev)	**Template-based (mean±stdev)**
Understanding of the description of the search task	3.63±0.56	3.47±0.63
Familiarity with the search topic before the start of the search session	2.53±1.04	2.47±1.14
Familiarity with the search topic after reading three related articles	3.3±0.59	3.13±0.63
Easiness in formulating the query	3.23±0.73	2.87*±0.78
Satisfaction with the search results	2.77±0.9	3.17±0.79
Overall satisfaction with the search system	2.73±0.87	3.1±0.71

As we can observe from the table, the users on average had similar levels of understanding of the descriptions of the search topics given to them and also they had similar average levels of familiarity with the assigned topic before the start of the search sessions. After showing them 3 sample relevant documents, the average levels of familiarity with the search topics increased from 2.53 to 3.3 and from 2.47 to 3.13 for control system and template-based system respectively.

The users found the process of formulating the query using templates (2.87) not as easy as using the single textbox in the control interface (3.23). The difference is statistically significant. One of the users explained in his/her response to the open-ended question: “It is not easy to put words in the query templates to fit the things I have in my mind to query.” Another user said: “The fixed patterns seem to add more difficulties in expressing the original meaning of the question”. Another user commented: “Having a single line of text instead of an entire text area felt restrictive. I also felt that the provided ‘questions’ weren't geared very well towards this topic, so I didn't use very many of them. However, I felt that the results returned by this search were much more relevant than in the previous condition.” Some users were concerned that they entered the same word into several templates, thus one of the users mentioned “trying to come up with new terms to avoid repetition” as one of the perceived difficulties.

The reported user satisfaction with the search results returned by the template-based system is on average 14.5% higher than that for the control system, and the overall satisfaction with the template-based search system is 13.41% higher than the satisfaction with the control system, although the differences are not statistically significant. One user said that the template-based system allowed him/her to “…express the type of information explicitly…” at the same time what this user disliked is that his/her information need required more than one template to express it. Another user commented, “The 13 lines in the query templates pretty much cover all question one would have, but it takes time to match my questions to the write line. Overall I feel the query templates are reasonably useful”. Some users felt that the terms they entered depended on the templates, e.g., “I found that I tend to come up with my search terms based on the question part in query templates, instead of independently creating the search terms”.

Chapter 6 Discussion

To achieve a better understanding of users' interaction with the templates, we performed a qualitative analysis of how users performed with given topics. As our user study goal required, we specifically analyzed query formulation related results.

6.1 User Interaction Analysis

In this section, we analyzed one group of six users who did the user study for same two topics: topic 406 and topic 443 with the different sequence of using control and experimental interfaces. In the user study design shown in table 6.1, 1 means the user used the corresponding system as the first interface and 2 means that he/she used it as the second interface.

Table 6.1 User study design for topic 406 and 443

User ID	Control System		Experimental System	
	Topic 406	Topic 443	Topic 406	Topic 443
User 01	1			2
User 06	2			1
User 11		1	2	
User 16		2	1	
User 21	1			2
User 30		2	1	

6.1.1 Topic 406

Topic 406, which is described as shown in table 6.2, was searched by 6 users. User01, user 06, and user21 were assigned this topic in their control system session; user11, user16, and user30 were assigned this topic in their template-based system session. Their search performance and logs were shown at Table 6.3 and Table 6.4 as below.

Table 6.2 Topic 406

Topic ID: 406

Topic Title: The diamond Industry

Description: What levels of corruption exist in the Diamond Industry and how does the development of a synthetic flawless diamond impact the industry?

Narrative: The diamond industry has long been associated with corruption on many levels. With the emergence of a new synthetic stone, the industry finds itself at a crossroads. Virtually anything dealing with corruption in the diamond industry and the effects generating from the development of the synthetic gem is on topic. Articles relating to the technical aspects of cutting and setting diamonds are off topic. Scientific articles are also off topic.

Table 6.3 Performance of the queries formulated for topic 406

Control system				**Experimental system**			
User	P@5	P@10	P@15	User	P@5	P@10	P@15
1	0.4	0.4	0.2667	11	0.8	0.6	0.6
6	0.2	0.4	0.2667	16	1	0.6	0.7333
21	0	0	0	30	1	0.8	0.5333
Average	0.2	0.2667	0.1778	Average	0.9333	0.6667	0.6222

Table 6.4 The number of unique query terms and average idf for topic 406

Control system			**Experimental system**		
User	Unique query terms excluding stopwords	Average idf of non-stopword query terms	User	Unique query terms excluding stopwords	Average idf of non-stopword query terms
1	4	5.1675	11	11	3.9805
6	18	5.2994	16	14	4.3694
21	6	5.116	30	8	3.1665
Average	9.3333	5.1943	Average	11	3.8388

As can be seen from Table 6.3 and 6.4, the three queries formulated using the experimental interface on average performed better in all of P@5, P@10, and P@15 since user21 judged all returned documents as not relevant when using control system. This user stated that he/she was not familiar with the topic at the beginning and he/she commented after the control system session that, “my strategy of adding a lot of words related to the topic did not pay off”. Even without considering this user’s results, the performance of the other two users was still much lower than the average level of using control interface. However, the performance of using experimental system by three users was much higher than the average level.

Although the average number of unique query terms entered via the template-based interface is higher than in the control, the average idf value in the template-based interface is much less than in the control. As can be seen from table 6.3, 11 unique words were entered in template-based interface with average idf value of 3.84, but the 9.33 unique words with average idf of 5.19 in the control. It is obvious that the experimental system with more words entered performed much better than the control even discarding the result of no relevant documents returned, as judged by user21.

Table 6.5 Filled in template by user 11 for topic 406

1. Find cases/instances of **african countries where blood diamond trading takes place.**
4. What effect does **"blood diamond"** have on **african war financing**?
5. What has been done (could be done) to alleviate/reduce the effect of **blood diamond** on **africa war finance?**
9. What impedes (gets in the way of) **regulating blood diamond**?
11. What methods/procedures are (can be) used to **regulate blood diamond**?

Table 6.6 Filled in template by user 30 for topic 406

1. Find cases/instances of **countries that has exported blood diamonds or conflict diamond.**
7. What do(es) **diamond companies** think/ say about **blood diamond or conflict diamond**?
13. Find statistical data on **wars and conflicts caused by blood diamond or conflict diamond?**

Table 6.7 Filled in template by user 16 for topic 406

1. Find cases/instances of **illegal diamond trade**.
2. What causes **blood diamond**?
4. What effect does **war in Africa** have on **illegal diamond trade**?
5. What has been done (could be done) to alleviate/reduce the effect of **illegal smuggling of diamond** on ________?
6. What has been done (could be done) to increase the effectiveness/effect of **Kimberley Process** on__?
7. What do(es) **Cilliers** think/ say about **Kimberley Process**?
8. Find positive and/or negative aspects (pros/cons) of **Resolution 1459**?
9. What impedes (gets in the way of) **Kimberley Process** ?
11. What methods/procedures are (can be) used to **stop illegal diamond trading**?
13. Find statistical data on **Countries who have illegal diamond trading** ?

Table 6.8 Queries entered by users using single textbox interface for topic 406

User01	diamond corruption blood synthetic
User06	The Kimberley Process Synthetic Diamonds Apollo Diamonds Created Diamonds Diamond Value Chain Corruption in Diamond Trading Corruption in Diamond Mining and Exploration Acceptance of
User21	bood diamonds corruption war synthetic flawless fake

For comparison, we listed all templates filled in and the queries formulated by these 6 users for the same topic 406 above. It is worth mentioning that user21 formulated the query with a typo word “bood” which should be “blood”. Most likely, this typo would be one of reasons of 0 relevant documents returned. For those three users who filled in the template-based questions to formulate queries, it is interesting to notice that there are many overlaps between the top 15 documents retrieved by the six users: 7 documents were retrieved twice and 1 document was retrieved by all three users. All such documents were judged consistently by all users. Moreover, as can be seen, the number of templates completed varies: 3 (user30), 5 (user11), 10 (user16). This example explained that the templates prompted the users to think about different aspects of the same topic based on their own knowledge background and information needs. For some templates, users entered very similar words and phrases, for example in template 1 “blood diamond trading”, “illegal diamond trade”, and “exported blood diamond”; for some other templates, users completed them quite differently, for example, user16 filled in “Kimberley Process” for template 6, 7, and 9, but the other users did not mention this at all. It is evident that the users filled in the templates based on their interests in different aspects of the topic.

6.1.2 Topic 443

Same as with topic 406, six users were involved in searching based on topic 443, which is described below. User 1, user6, and user 21 searched this topic by using the templates interface, and user11, user16, and user30 searched this topic by using the control interface.

Table 6.9 Topic 443

Number: 443

Title: Hand-Held Electronics

Description: New technologies in the areas of cell phones, PDAs, and so forth.

Narrative: As technology more and more focuses on adapting old technology for smaller scales, new advances are abounding in the consumer markets, including advances in cell phone technology, PDA technology, notepad PCs, and digital cameras. Advances in all types of hand-held technology are on topic. Reports on large-scale technology, unless substantially applicable to hand-held technology as well are off topic. In such a case, the article itself must make the cross-application to hand-held technology.

Table 6.10 Filled in templates by user 1 for topic 443

1. Find cases/instances of **portable electronics.**
3. What are existing/potential uses/applications of **portability** (to/for/in **electronics**)?
6. What has been (or could be) done to increase the effectiveness of **miniaturization** on **electronics**?
8. Find positive and/or negative aspects (pros/cons) of **hand-held devices**?
10. What is the goal/purpose of **portability**?
11. What methods/procedures are used (can be used) to do/achieve **make devices smaller**?
13. Find statistical data on **portable electronics**?

Table 6.11 Filled in templates by user 6 for topic 443

1. Find cases/instances of **mobile phone, PDA, digital cameras.**
2. What causes **advancement in technology?**
4. What effect does **internet** have on **mobile phone**?
5. What has been (or could be) done to alleviate/reduce the effect of **website compatibility** on **PDA**?
8. Find positive and/or negative aspects (pros/cons) of **digital vs. film**?
9. What impedes (gets in the way of) **trading with PDA**?

Table 6.12 Filled in templates by user 21 for topic 443

1. Find cases/instances of **hand-held electronics.**
2. What causes **new hand-held technologies to succeed?**
3. What are existing/potential uses/applications of **smart phones** (to/for/in _____)?
4. What effect does **UMTS** have on **cell phone adaption rates**?
5. What has been (or could be) done to alleviate/reduce the effect of **location** on **poor reception**?
6. What has been (or could be) done to increase the effectiveness of **messaging** on **cellphones**?
7. What do(es) **RIM** think about **the blackberry**?
8. Find positive and/or negative aspects (pros/cons) of **advances in hand-held technology**?
9. What impedes (gets in the way of) **consumer awareness of new cellphone features**?
10. What is the goal/purpose of **PDA innovation**?
11. What methods/procedures are used (can be used) to do/achieve **link multiple wireless devices**?
12. Compare **iPhone** and **Blackberry** in terms of ______.
13. Find statistical data on **smart phone adaption rates**?

As can be seen from Tables 6.10,11, and 12, the templates-based interface encouraged users to fill in many templates when articulating their information needs, for example user 21 filled in all 13 templates. From this example, it is easily observed that the templates prompted the users to think about different aspects of cell phone technology and industry, such as the effect of UMTS (Universal Mobile Telecommunications System) on cell phone adoption rates, methods of linking multiple wireless devices and comparison of iPhone and Blackberry. Similar to topic 406, for some templates the users entered very similar words and phrases, for example in template 1: “portable electronics”, “hand-held electronics” and “mobile phone, PDAs, digital cameras”, however for some other templates users entered quite different queries, for example, template 4 was completed as “What effect does **internet** have on **mobile phone**?” by user 6, and as “What effect does **UMTS** have on **cell phone adaption rates**?” by user 21. Another example is template 11: user 1 filled it in as “What methods/procedures are used (can be used) to do/achieve **make devices smaller**?”, while user 21 completed it as “What methods/procedures are used (can be used) to do/achieve **link multiple wireless devices**?”. It is this difference that the users filled in the templates based on their knowledge of and interest in different aspects of the topic can help to explain the overall low agreement in filling in the templates for the same topic [32].

The users’ performance and related log records for this topic are shown in table 6.13 and table 6.14 below. As can be seen, the three queries formulated using the experimental interface on average performed better in P@5 and P@10. One more time, the average number of unique query terms entered via the template-based interface is higher than in the control. However, more words do not always lead to higher performance. As evident from Table 6.14, users 21 and 6 entered 27 and 12 unique words respectively (with the average *idf* of 5.114 and 4.964), but the P@15 was 0.4 and 0.667 respectively. On the other hand, the overlap between the top 15 documents retrieved by the six users given topic 443 is low: 74 out of 90 documents were retrieved only once, one document was retrieved by four users, and six documents were retrieved by two users, but only one document was judged inconsistently by two users.

Table 6.13 Performance of the queries formulated using the template-based and control interfaces for topic 443

Experimental system				Control system			
User	P@5	P@10	P@15	User	P@5	P@10	P@15
1	0.6	0.4	0.267	11	0.6	0.4	0.267
6	1	0.9	0.667	16	1	0.7	0.733
21	0.6	0.4	0.4	30	0.2	0.3	0.333
Average	0.73	0.56	0.445	Average	0.6	0.46	0.444

Table 6.14 Number of unique query terms and average *idf* in the queries formulated for topic 443

Experimental system			Control system		
User	Unique query terms excluding stopwords	Average idf of non-stopword query terms	User	Unique query terms excluding stopwords	Average idf of non-stopword query terms
1	6	6.256	11	7	6.053
6	12	4.964	16	8	4.712
21	27	5.114	30	12	4.045
Average	15	5.445	Average	9	4.937

6.1.3 Topic 409

Another interesting topic is topic 409 (Table 6.15), which was assigned to three users using the template-based system (user2, user7 and user 22) and three users using the control system (user12, user17 and user26). Below we list all the user-entered queries and analyze the system performance.

As can be seen from table 6.16, both systems performed worse than that the average level performance in P@5, P@10, and P@15. Compared with the template-based system, the control system performed much better.

Table 6.15 Topic 409

Number: 409

Title: AIDS in Africa

Description: What is the state of AIDS in Africa?

Narrative: Little attention has been given to the AIDS epidemic in Africa that has decimated an entire generation of Africans. What is being done to help prevent the spread of AIDS and to treat those already infected? What sorts of public education/health measures have African governments taken? What are the barriers?

Table 6.16 Performance of topic 409

Control system				**Experimental system**			
User	P@5	P@10	P@15	User	P@5	P@10	P@15
12	0.4	0.3	0.333	2	0.4	0.3	0.333
17	0.8	0.8	0.867	7	0.4	0.5	0.4
26	0	0	0.2	22	0	0.2	0.2
Average	0.4	0.3667	0.4667	Average	0.2667	0.3333	0.3111

As user 22 commented about the templates “some of the questions do not apply to the topic, and some of them do. This makes me a little confused”, it is evident that users felt the process of formulating the query using templates not as easy as using single textbox in control interface. Another factor which influenced the performance and user query input strategies for the templates-based interface is that users entered the same word into several templates, as user 7 stated that “an information need could be expressed through multiple templates” and also another user commented his/her query entering strategy as “trying to come up with new terms to avoid repetition”. Tables 6.17, 6.18, 6.19, 6.20 list the users’ queries entered using the templates and the control system respectively.

Table 6.17 Filled in templates by User 2 for topic 409

1. Find cases/instances of **AIDS epidemic Africa**.
2. What causes **AIDS**?
3. What are existing/potential uses/applications of **drugs** (to/for/in **AIDS**)?
4. What effect does **drugs** have on **AIDS**?
5. What has been done (could be done) to alleviate/reduce the effect of **AIDS** on **Africa** ?
6. What has been done (could be done) to increase the effectiveness/effect of **drugs** on **AIDS** ?
7. What do(es) **African govenments** think/ say about **AIDS epidemic** ?
8. Find positive and/or negative aspects (pros/cons) of **AIDS**?
9. What impedes (gets in the way of) **AIDS Prevention**?
11. What methods/procedures are (can be) used to **prevent AIDS** ?
13. Find statistical data on **AIDS Africa**?

Table 6.18 Filled in templates by User 7 for topic 409

1. Find cases/instances of **measures taken by African Governments to prevent AIDS** .
4. What effect does **war(s)** have on **public Health education (especially AIDS related issues)**?
9. What impedes (gets in the way of) **preventing the spread of (and treating) AIDS** ?
13. Find statistical data on **success of AIDS prevention measures**?

Table 6.19 Filled in templates by User 22 for topic 409

1. Find cases/instances of **aids prevention** .
3. What are existing/potential uses/applications of **treat aids** (to/for/in____)?
4. What effect does **public education/health measures** have on **AIDS**?
5. What has been done (could be done) to alleviate/reduce the effect of **spread of aids** on_?
6. What has been done (could be done) to increase the effectiveness/effect of **prevent the spread of aids** on _____?
7. What do(es) **goverment**__ think/ say about **aids prevention**?
11. What methods/procedures are (can be) used to **prevent aids from spread**?

Table 6.20 Formulated queries by using control interface for topic 409

User 12	Africa AIDS HIV action prevention treatment education health help government difficulty problem barrier
User 17	HIV AIDS anti-retroviral treatment therapy protest public sector private sector
User 26	aids prevent education health measures barriers africa government vaccine anti viral

As can be seen from above tables, the users filled in the templates based on their requirements: user 2 filled 11, user 7 filled 4, and user 22 filled 7 templates. All three users answered question 1 “Elaboration” and 4 “Result”. The users’ interest differences also can help to explain this overall low agreement in filling the templates for the same topic. On the other hand, it is evident that users’ knowledge background led to different aspects of the topic, which can be represented by means of these templates. For example, for question 4, user 2 wanted to know the AIDS drug effect, user 7 was interested in the war(s) effect on AIDS related issues, and user 22 cared about the effect of public education/health measures on AIDS. It is worth mentioning that for this topic users entered same words many times, for example for every answered template “AIDS” was entered and also some other words like “drug”, “Africa”, and “Africa” have been used many times. However, there is very little overlap between the returned documents by the six users: 23 out of 90 documents were retrieved more than once, of which 10 documents were retrieved by two users and one document retrieved by three users. There is no inconsistent judgment for these documents by users.

6.2 Interfaces Discussion

Belkin, et al. [36] compared two query-elicitation modes, a query-entry line and a scrollable query-entry box, to determine whether box mode (which allowed searchers to enter and see a complete query of five 40 character- long lines at a time) would result in longer queries than the line mode (which allowed searchers to see 50 characters of a query). They found that the box mode led to somewhat longer queries, and that the full sentence/question type led to significantly longer queries [36]. Although their study was not explicitly designed to evaluate the effect of query length on performance, they did find a consistent relationship between query length and one measure of performance in the task.

In general, research in this area has positively proven few hypotheses: a search interface which asks searchers to describe their information needs at length leads to longer queries than the one which asks searchers to simply input a query as a list of words or phrases; query length is positively correlated with performance in the search task; a system which encourages long queries leads to better performance in the search task [2, 37, 38, 39].

While in operational IR settings, the average query length is 2-3 words [39], in our experiment both interfaces encouraged users to enter more queries. As we discussed in chapter 5, the average number of unique query terms excluding stopwords entered using both control and template systems is 11.12±5.42. In the template-based interface, the number of unique query terms without stopwords is higher than in the single-textbox interface and the performance of template-based interface is better than in control interface. But as we discussed above, for some specific topics, more words do not necessarily mean better performance.

Chapter 7 Conclusions and future work

For some users who have complex information needs, their need was defined in our understanding as a need to learn about specific semantic relations between entities or aspects of an entity. In this study, an approach was proposed to help users articulate complex information need by filling in topic-independent query templates. We asked users to fill in a list of templates, formulated as questions or statements based on the Rhetorical Structure Theory (RST). We hypothesized that such templates encouraged users to think about and express their information needs in terms of relations between the entities of their interest, such as what causes X, what can be done to prevent X, what effect does X have on other entities, and find statistical data on X, etc. Motivated by finding answers to our two research questions shown below, we designed a control system with a single query textbox and an experimental interface with the query templates, and conducted a user study with 30 users.

RQ1: Are users able to articulate complex information needs using topic-independent query templates?

The data gathered based on query records and search session logs suggest that users are able to express their information needs by filling in query templates. On average the users completed 7.8 templates out of 13. Also, the average number of unique query terms excluding stopwords entered using the template-based system is slightly higher than that of the control system.

RQ2: When used with a bag-of-words retrieval system, do the queries formed by using template-based interface lead to better search results than queries formed by using a single-textbox interface?

The template-based interface helped the users formulate on average more effective queries, resulting in better search results (measured in precision at 5, 10 and 15 documents) compared to the control system. Improvement in P@10 is statistically significant. The average user satisfaction levels with the search results and the overall system reported in the questionnaire are also higher for the template-based system. At the same time, the users' subjective ratings of the easiness of formulating queries indicate that on average they found query formulation by filling in templates not as easy as entering their query into one textbox. However, the average time spent on formulating the query using templates is only 27 seconds longer than using one textbox.

Some of the users found some templates to be too restrictive. Such templates need to be re-written to make them clearer and relevant to a wider range of queries. For example, one user commented that one of the templates "is a little hard to understand. Maybe some re-wording will help". Another user mentioned that "Having a single line of text instead of an entire text area felt restrictive".

In the present work our goal was to study template-filling as a query formulation method, we did not use the relationships specified by the users in the query templates in document retrieval and ranking. An interesting future research direction may be about automatic identification of semantic relations in text. Research in this direction is done, for example, within the framework of RTE (Recognizing Textual Entailment) track of TAC (Text Analysis Conference) [43]. Such methods may facilitate the retrieval of documents to match the specific relations expressed by the user in the query. Moreover, the research aimed at how to improve the templates and make them more related to users' specific information needs may be another interesting direction. For example, a system which could automatically select a subset of templates to show to the user on the basis of his/her initial short query statement of his/her information needs, would be desirable by tuning the templates and avoiding showing to the user those templates that are unlikely to be relevant. In this way it would be possible to build a larger set of detailed templates, with only some of them being shown to the user as fill-in-the-blanks clarification questions.

References

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

2 Kelly D., Dollu V. J. and Fu X. The loquacious user: a document-independent source of terms for query expansion. 28th ACM-SIGIR, Bahia, Brazil, 2005.

3 Kelly, D. and Lin, J., Overview of the TREC 2006 ciQA task. ACM-SIGIR Forum, Vol.41 , Issue 1, pp. 107 – 116. 2007

4 Belkin N. J., Oddy R. N. and Brooks H. M. ASK for Information Retrieval: Part I. Background and Theory. Journal of Documentation, 38(2), pp. 61-71, 1982.

5 Belkin N. J., Cool C., Kelly D., Lin S., Park S. Y., Perez-Carballo J. and Sikora C. Iterative exploration, design and evaluation of support for query reformulation in interactive information retrieval. Information Processing and Management, 37(3), 404-434.

6 Jansen B.J., Spink A. and Saracevic T. Real life, real users and real needs: a study and analysis of user queries on the web. Information Processing and Management, 26, pp. 207-227, 2000.

7 Baeza-Yates, R., Riberio-Neto, B., Modern Information Retrieval. Chapter 2: Modeling, chapter 3: Retrieval Evaluation. 1999

8 Robertson, S.E., Walker, S. and Hancock-Beaulieu, M.M., Experimentation as a way of life: Okapi at TREC. Information Processing and Management 36(1), pp. 95-108, 1999.

9 Chiaramella, Y., and Chevallet, J. P. About retrieval models and logic. The computer journal, 35(3), 233-242, 1992.

10 Brin S., and Page L. The Anatomy of a Large-Scale Hypertextual Web Search Engine. Computer Networks and ISDN Systems, Volume 30 , Issue 1-7, pp 107

– 117, 1998.

11 Robertson, S. E. and Sparck Jones, K. Relevance weighting of search terms. Journal of the American Society for Information Science, 27, 1976.

12 Sparck Jones, K., Walker, S. and Robertson, S. E. A probabilistic model of information retrieval: development and comparative experiments. Information processing and management, 36(6), 2000.

13 Kagolovsky, Y. and Moehr, J. Current Status of the Evaluation of Information Retrieval. Journal of Medical Systems, Vol. 27, No. 5, October 2003.

14 Belkin, N., Scholtz, J., Dumais, S. and Wilkinson, R. Evaluating Interactive Information Retrieval Systems: Opportunities and Challenges. CHI 2004. 24-29 April Vienna, Austria.

15 Voorhees, E. M., Buckley, C. Retrieval Evaluation with Incomplete Information. SIGIR Conference 2004 (Sheffield, UK, 2004).

16 David Hull. Using Statistical Testing in the Evaluation of Retrieval Experiments. SIGIR 1993 Pittsburgh, PA, USA.

17 Voorhees, E. Overview of TREC 2006. http://trec.nist.gov/pubs/trec15/papers /OVERVIEW.pdf, National Institute of Standards and Technology, Gaithersburg, MD 20899.

18 Spark Jones, K. , Walker, S., Gatford, M., and Do, T. Peeling the onion: Okapi system architecture and software design issues. Journal of Documentation, Volume 53, Issue 1, pp 58 - 68, 1997.

19 Okapi-Pack. http://www.soi.city.ac.uk/~andym/OKAPI-PACK/index.html (last visited on 08/12/2008), Centre For Interactive Systems Research, City University, London EC1V 0BH

20 Robertson, S. Development of IR evaluation methods: Okapi at TREC. 1999

21 Robertson, S. E.; Walker, S. and Beaulieu, M. Experimentation as a way of life:

Okapi at TREC. Information Processing & Management, 36(1), 95-108. 2000

22 Vechtomova O., Karamuftuoglu M. (2006) Elicitation and Use of Relevance Feedback Information. Information Processing and Management, Vol.42, Issue 1, pp. 191-206

23 Efthimiadis, E. N. Query expansion. Annual Review of Information Science & Technology, 31. 1996

24 Rocchio, J. Relevance feedback in information retrieval. The SMART Retrieval System: Experiments in Automatic Document Processing. 1971.

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

26 Sakai, T., Manabe T., and Koyama M. Flexible Pseudo-Relevance Feedback via Selective Sampling. ACM Transactions on Asian Language Information Processing, Vol. 4, No. 2, June 2005, Pages 111–135.

27 Ruthven, I. Re-examining the potential effectiveness of interactive query expansion. Proceeded in the 26th Annual ACM International Conference on Research and Development in Information Retrieval (SIGIR '03), Toronto, CA, 213-220. 2003

28 Taboada, M. and Mann1, W. Applications of Rhetorical Structure Theory. July 22, 2005.

29 http://www.sfu.ca/rst/01intro/definitions.html (Last visited on 08/12/2008)

30 Mann, W.C., and Thompson, S.A. Rhetorical Structure Theory: Toward a functional theory of text organization. Text, 8 (3). 243-281. 1988

31 http://trec.nist.gov/. (Last visited on 08/12/2008)

32 Vechtomova, O., Zhang, H. Articulating complex information needs using Query Templates. Journal of Information Science (Accepted: December, 2008).

33 Spärck Jones K, Walker S. and Robertson S.E. A probabilistic model of information retrieval: development and comparative experiments. Parts 1 and 2. Information Processing and Management, 2000, 36, pp. 779-808, 809-840.

34 Allan, J. HARD track overview in TREC 2004. High accuracy retrieval from documents. In E. Voorhees & L. Buckland (Eds.), Proceedings of the 13th text retrieval conference, Gaithersburg, MD, USA.

35 http://www.readingsoft.com. (Last visited on 08/12/2008)

36 Belkin, N.J., Cool, C., Kelly, D., Kim, G., Kim, J.-Y., Lee, H.-J., Muresan, G., Tang, M.-C., & Yuan, X.-J. (2003). Rutgers Interactive Track at TREC 2002. In E.M. Voorhees& D. M. Harman (Eds.). The 2002 Text Retrieval Conference.

37 Ruthven, I. Re-examining the potential effectiveness of interactive query expansion. Proceeded in the 26th Annual ACM International Conference on Research and Development in Information Retrieval (SIGIR ’03), Toronto, CA, 213-220. 2003

38 Belkin, N. J., Cool, C., Kelly, D., Lee, H.-J., Muresan, G., Tang, M.-C., & Yuan, X.-J. (2003). Query length in interactive information retrieval. In Proceedings of the 26th Annual ACM International Conference on Research and Development in Information Retrieval (SIGIR ’03), Toronto, CA, 205-212.

39 Jansen B.J., Spink A. and Saracevic T. Real life, real users and real needs: a study and analysis of user queries on the web. Information Processing and Management, 26, pp. 207-227, 2000.

40 Vechtomova O. Articulating complex information needs using query templates. Unpublished working paper, 2007.

41 Thompson D. Interface Design for An Interactive Information Retrieval System: A literature survey and a research system description. Journal of the American Society for Information Science, Volume 22 Issue 6, Pages 361 – 373, 2007.

42 Jones S., Vechtomova O., Robertson S. A tool for comparative evaluation in an interactive environment. Journal of Information Science, 28(6), 2002.

43 Giampiccolo D. and Trang Dang H. (track coordinators) TAC 2008 Recognizing Textual Entailment (RTE) Track (http://www.nist.gov/tac/tracks/2008/rte/)

Appendices

Appendix A: Recruitment Email

Fellow Graduate Students:

My name is Hao Zhang and I am a graduate student in the department of Management Sciences. I am currently doing my thesis in the Information Retrieval (IR) area supervised by Prof. Olga Vechtomova. The goal of this study is to see whether a query template can be useful in helping users with complex information needs by articulating and eliciting more search query terms. This research will hopefully lead to a better understanding of what should be done to improve the quality of the template in terms of high performance of the IR system.

If you volunteer as a participant in this study, you will be asked to logon to a designed experimental Information Retrieval system from the computer in office CPH4332 and go through two search tasks related to the given topics. You will also need to fill in the questionnaire form after every search task. The whole session should take approximately two hours of your time. In appreciation of your time, you will receive $20.

I would like to assure you that this study has been reviewed and received ethics clearance through the Office of Research Ethics. However, the final decision about participation is yours.

If you are interested, please send me an e-mail at h13zhang@engmail.uwaterloo.ca and I will be in touch with you to arrange the best convenient time for you to participate in the research. Alternatively, you can come to my office CPH4332 and see me directly.

Hao Zhang

Management Science Department, University of Waterloo

Email: h13zhang@engmail.uwaterloo.ca

Appendix B: Informational Letter

Department of Management Sciences

University of Waterloo

Date: May 8th, 2008

Title of Project: The Use of Query Templates in Complex Interactive Question Answering (ciQA)

Principal Investigator: ***Prof. Olga Vechtomova***
Management Sciences Department, University of Waterloo
(519) 884-4567 Ext. 32675

Student Investigator: ***Hao Zhang***
Management Sciences Department, University of Waterloo
(519) 884-4567 Ext. 36005

The goal of this study is to see whether a set of domain-independent query templates can be useful in helping the user to articulate their information need, and to elicit more query terms from the user. The findings from this study will be important in advancing our understanding of how to design user interface to solve complex information needs in an Information Retrieval (IR) system. These findings have potential application in improving expressive query language that can be used in helping users to formulate detailed complex queries.

Twenty people in total will be asked to participant this research. As a participant in this study, you will be assigned two search tasks. Before you logon the system, you will need to fill in an entry questionnaire form which will collect background information about you. At the beginning of each task, the search topic and the description of the information requirement will be given. Then, the IR system will ask you to read the 3 documents judged as relevant to help you understand the topic and the complex information needs. One task will be done by using the control system which allows you to input as many of the search queries as you want into a text area. Based on what you put as the search queries, the system will return the top 15 documents. You will read them one by one to judge it as relevant or not. The other task will be done by using the experimental system which asks you to answer 13 given questions by filling in the blanks. You are free to input your answers into these questions and you may also skip the questions you may think that are not related with the topic and information requirement description. Also, the system will return top 15 documents and ask you to do the relevance judgment. At the end of each task, you will be asked to fill in a feedback questionnaire. You may decline to answer questions if you wish. In total, the study will require about two hours of your time.

The collected data will be coded with participant numbers (not names) and will be kept indefinitely in a secure place as electronic format. You may withdraw from the study at any time without penalty by verbally indicating this to the researcher. There are no known risks associated with participating in this study. In appreciation for your time you will receive $20. If you withdraw from participation, remuneration will be pro-rated at $7.50/hour.

I would like to assure you that this study has been reviewed and received ethics clearance through the Office of Research Ethics at the University of Waterloo. However, the final decision about participation is yours. Should you have any comments or concerns resulting from your involvement in this study, please contact Dr. Susan Sykes in the Office of Research Ethics at 519-888-4567 Ext. 36005.

If you have any questions later or require additional information about the study, please feel free to contact either of the researchers at (519) 884-4567 Ext. 32567 or at (519) 888-4567 Ext. 36005.

CONSENT FORM

I agree to take part in a research study being conducted by Dr. Olga Vechtomova and graduate student Hao Zhang of the Department of Management Sciences, University of Waterloo.

I have made this decision based on the information I have read in the Information letter. All the procedures, any risks and benefits have been explained to me. I have had the opportunity to ask any questions and to receive any additional details I wanted about the study. If I have questions later about the study, I can ask one of the researchers above.

I understand that I may withdraw from the study at any time without penalty by telling the researcher.

This project has been reviewed by, and received ethics clearance through, the Office of Research Ethics at the University of Waterloo. I am aware that I may contact this office (519-888-4567, ext. 36005) if I have any concerns or questions resulting from my involvement in this study.

____________________________ ____________________________

Printed Name of Participant Signature of Participant

Dated at Waterloo, Ontario Witnessed

Appendix C: Entry Questionnaire before using IR system

Q1: In what age group are you?

- 19 and under
- 20 – 29
- 30 – 39
- 40 – 49
- 50 - 59

Q2: Gender:

- Male
- Female

Q3: How often do you use search engines?

- Every day
- Several times per week
- Less than once a week
- Very rarely

Q4: What do you use search engines mostly for?

- Work/Study
- Personal use
- Both

Q5: Do you use advanced search functions of search engines?

- Yes
- No

Appendix D: Feedback Page

Dear Participant,

We would like to thank you for your time and commitment to this study. It is very valuable to our research.

The data collected during study will be used to improve the performance of future Information Retrieval Systems. I would like to assure you that all information you provided will be kept confidential; your name will not appear in any data records or publications.

A summary of the main findings of the project will be made available to all participants upon request. Please email: h13zhang@engmail.uwaterloo.ca if you are interested in receiving this information as soon as the research is completed. This project was reviewed by, and received ethics clearance through, the Office of Research Ethics at the University of Waterloo. Should you have any comments or concerns resulting from your participation in this study, please contact Dr. Susan Sykes in the Office of Research Ethics at 519-888-4567, Ext., 36005.

Sincerely,

Hao Zhang
Management Sciences Department
University of Waterloo
Email: h13zhang@engmail. uwaterloo.ca

Appendix E: ORE letter

Dear Researcher:

A Request for ethics review of a modification or admendment (ORE 104) to your ORE application:

Title: The Use of Query Templates in Interactive Complex Question Answering
ORE #: 14634
Faculty Supervisor: Olga Vechtomova (ovechtom@engmail.uwaterloo.ca)
Student Investigator: Hao Zhang (h13zhang@engmail.uwaterloo.ca)

The proposed modification request has been reviewed and has received full ethics clearance.

A signed copy of the 'Request for Ethics Clearance of a Modification to an Ongoing Application to Conduct Research with Human Participants' will be provided through regular mail. In the case of student research, the signed copy will be sent to the Faculty Supervisor.

Note 1: This project must be conducted in accordance with the description in the application and modification for which ethics clearance has been granted. All subsequent modifications to the protocol must receive prior ethics clearance through the Office of Research Ethics.

Note 2: Researchers must submit a Progress Report on Continuing Human Research Projects (ORE Form 105) annually for all ongoing research projects. In addition, researchers must submit a Form 105 at the conclusion of the project if it continues for less than a year.

Note 3: Any events related to the procedures used that adversely affect participants must be reported immediately to the ORE using ORE Form 106.

Susanne Santi, M. Math.,
Manager
Office of Research Ethics
NH 1027
519.888.4567 x 37163
ssanti@uwaterloo.ca

Appendix F: Questionnaire after using the control interface

Q1: How well did you understand the description of the search task?

- Very well
- Well
- Not too well
- Did not understand at all.

Q2: How familiar were you with the search topic before the start of today's session?

- Very familiar
- Familiar
- Not too familiar
- Not familiar at all

Q3: How familiar were you with the search topic after reading three related articles given to you at the beginning of the search session?

- Very familiar
- Familiar
- Not too familiar
- Not familiar at all

Q4: How easy was it for you to formulate the query in your own words?

- Very easy
- Moderately easy
- Somewhat difficult

- Very difficult

Q5: Please describe difficulties (if any) in formulating the query:

____

____

____

Q6: How satisfied are you with the search results?

- Very satisfied. I found most of the information I was searching for.
- Somewhat satisfied. I found some relevant information.
- Not too satisfied. Most of the sought information was not found.
- Not satisfied at all. None of the sought information was found.

Q7: Please rate your overall satisfaction with this search system.

- Very satisfied.
- Somewhat satisfied.
- Not too satisfied.
- Not satisfied at all.

Appendix G: Questionnaire after using the experimental interface

Q1: How well did you understand the description of the search task?

- Very well
- Well
- Not too well
- Did not understand at all.

Q2: How familiar were you with the search topic before the start of today's session?

- Very familiar
- Familiar
- Not too familiar
- Not familiar at all

Q3: How familiar were you with the search topic after reading three related articles given to you at the beginning of the search session?

- Very familiar
- Familiar
- Not too familiar
- Not familiar at all

Q4: How easy was it for you to formulate the query using the query templates?

- Very easy
- Moderately easy

- Somewhat difficult
- Very difficult

Q5: Please describe difficulties (if any) in formulating the query using the query templates:

____

____

____

____

____

Q6: How satisfied are you with the search results?

- Very satisfied. I found most of the information I was searching for.
- Somewhat satisfied. I found some relevant information.
- Not too satisfied. Most of the sought information was not found.
- Not satisfied at all. None of the sought information was found.

Q7: Please rate your overall satisfaction with this search system.

- Very satisfied.
- Somewhat satisfied.
- Not too satisfied.
- Not satisfied at all.

Appendix H: Titles of TREC Topic 301-450

Topic ID	Topic Title
301	Abuses of E-Mail
302	Adoptive Biological Parents
303	African Civilian Deaths
304	Agoraphobia
305	Airport Security
306	alternative medicine
307	Alzheimer's Drug Treatment
308	Amazon rain forest
309	anorexia nervosa bulimia
310	Antarctica exploration
311	Antibiotics Bacteria Disease
312	Argentine/British Relations
313	automobile recalls
314	Best Retirement Country
315	Black Bear Attacks
316	blood-alcohol fatalities
317	British Chunnel impact
318	cigar smoking
319	clothing sweatshops
320	commercial cyanide uses
321	Cult Lifestyles
322	Diplomatic Expulsion
323	dismantling Europe's arsenal
324	drug legalization benefits
325	Educational Standards
326	El Nino
327	encryption equipment export
328	Endangered Species (Mammals)
329	euro opposition
330	Export Controls Cryptography
331	Falkland petroleum exploration
332	Ferry Sinkings
333	food/drug laws
334	Health and Computer Terminals
335	health insurance holistic
336	home schooling
337	Hubble Telescope Achievements
338	Human smuggling

339	hybrid fuel cars
340	hydrogen energy
341	hydrogen fuel automobiles
342	Hydroponics
343	illegal technology transfer
344	Implant Dentistry
345	in vitro fertilization
346	Income Tax Evasion
347	Industrial Espionage
348	International Art Crime
349	International Organized Crime
350	Iran-Iraq Cooperation
351	journalist risks
352	Land Mine Ban
353	Literary/Journalistic Plagiarism
354	Magnetic Levitation-Maglev
355	mainstreaming
356	Marine Vegetation
357	Mental illness drugs
358	mercy killing
359	Metabolism
360	Mexican Air Pollution
361	Modern Slavery
362	Most Dangerous Vehicles
363	Mutual fund predictors
364	Native American casino
365	New Fuel Sources
366	New Hydroelectric Projects
367	Nobel prize winners
368	Obesity medical treatment
369	ocean remote sensing
370	oceanographic vessels
371	organic soil enhancement
372	Orphan drugs
373	Overseas Tobacco Sales
374	Piracy
375	Police Deaths
376	Poliomyelitis and Post-Polio
377	Polygamy Polyandry Polygyny
378	Pope Beatifications
379	postmenopausal estrogen Britain
380	R&D drug prices
381	Rabies

382	Radio Waves and Brain Cancer
383	radioactive waste
384	Rap and Crime
385	Risk of Aspirin
386	robotics
387	sick building syndrome
388	space station moon
389	teaching disabled children
390	territorial waters dispute
391	Tourism
392	transportation tunnel disasters
393	Undersea Fiber Optic Cable
394	Unexplained Highway Accidents
395	Unsolicited Faxes
396	Viral Hepatitis
397	Wildlife Extinction
398	Women in Parliaments
399	World Bank Criticism
400	World Court
401	Bass Amps
402	Identity Theft
403	Heaven's Gate
404	Marathon Training
405	Female competitive fighters
406	The Diamond Industry
407	Chimpanzee Language Ability
408	College Campus Racism
409	AIDS in Africa
410	Low-Carb Mania
411	Natural Disasters and Global Warming
412	Outsourcing
413	The Future of Corn
414	Human Evolution
415	Life on Mars
416	Blogging
417	Diabetes Research
418	Immigration Post 9-11
419	Do It Yourself Computer Building
420	Internet Security through Quantum Computing
421	Software Patents
422	Video game crash
423	United Nations Development Programme's Millennium Declaration
424	Bollywood

425	Cancun World Trade Talks
426	Chinese capitalism
427	Brazilian Landless Workers Movement
428	International organ traffickers
429	Biodynamic and Organic farming
430	Chlorofluorocarbon Ban
431	Britney Spears
432	Farm Subsidies
433	Sexally Transmitted Diseases Among Teens
434	The Zapatista movement
435	LeBron James
436	Dinosaurs
437	Role Playing Games
438	BP: Beyond Petroleum
439	Giant Squid
440	Supreme Court Rulings
441	European Union
442	Interest Rates
443	Hand-Held Electronics
444	European Elections
445	Bird Migration Deaths
446	Grand Canyon Environment
447	VX nerve gas disposal
448	NASA
449	Multiple Pregnancies
450	Fractals