JIS Accepted for Publication
By the Journal of Information Science: http://jis.sagepub.co.uk

Articulating Complex Information Needs Using Query Templates

Olga Vechtomova[1] and Hao Zhang

Department of Management Sciences, University of Waterloo

ABSTRACT

In this paper we investigate the effectiveness of topic-independent query templates as a tool for assisting users in articulating their information needs. We hypothesize that topic-independent query templates can help users with complex information needs to express their requirements more accurately and in greater detail. We developed a set of query templates representing general semantic relationships between concepts, such as cause-effect and problem-solution. Each template was written in the form of a fill-in-the-blanks question. A user study was performed comparing the template-based interface with a single-textbox search interface. Results demonstrate 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.

Keywords: Query formulation, query templates, complex information needs, Rhetorical Structure Theory, interactive information retrieval, user study.

1. Introduction

While many users have relatively simple or general information needs, users who are familiar with a certain topic may have more specific or complex information needs. Having already some knowledge of a subject and its concepts or entities, such 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, e.g., how does entity X

[1] *Correspondence to: Olga Vechtomova, Department of Management Sciences, University of Waterloo, 200 University Avenue West, Waterloo, ON, N2L 3G1, Canada, ovechtom@engmail.uwaterloo.ca.*

JIS-0832-v2 Received: 24th September 2008 Revised: 9th December 2008

influence entity Y or what do entities X and Y have in common. We refer to such information needs as complex needs. Consider the following scenario: a business analyst wants to find information about the effects of credit card fraud on the U.S. banks. She has been working in banking and finance for a long time, therefore, she is not interested in documents explaining the banking system or types of credit card fraud. Instead, she would like to learn about the financial losses that banks incur due to credit card fraud and whether a growing number of fraud cases and banks' insufficient fraud prevention programs cause customers to leave.

In contrast, users unfamiliar with a subject have limited or no knowledge of its concepts and entities, and hence are less likely to have a clear understanding of what they need to know in this subject area. Their information needs are, therefore, likely to be relatively simple: to find information *about* the subject, so that they can learn more about it. An example scenario is a first-year university student beginning to work on a coursework paper about credit card fraud. He has no previous knowledge of this subject and does not have a clear idea what he needs to know. So, first he would like to read basic information about the subject.

In order to resolve complex information needs, an IR system, should, firstly, help users express different semantic relations between the concepts/entities of their interest. Secondly, it should have a retrieval model that would rank documents based on the likelihood that they contain the sought semantic relations between entities.

In this paper we focus on the first problem: helping users articulate complex information needs by expressing specific semantic relations between the entities of their interest. 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-the-blanks questions. Each of the templates expresses a topic-independent semantic relation, such as cause-effect and problem-solution. The user is invited to fill in as many or as few templates as they feel necessary to express their information need. An example of a template is "What effect does ____ have on ____?"

We implemented two query formulation interfaces: a template-based interface and a single-textbox interface, and conducted a user study to learn how effectively users formulate queries with them. It is reasonable to expect that a template-based query formulation interface may be more effective as a front-end to an IR system that makes use of semantic relations in document ranking, as opposed to a bag-of-words IR system. However, no such system is presently available, and the developed interfaces were evaluated as front-ends to one of the state-of-the-art IR systems, Okapi [1]. The goal of the user study was to investigate whether users can be expected to effectively articulate complex information needs by filling in topic-independent templates. Two major research questions we explored are:

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

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?

The rest of the paper is organized as follows: Section 2 discusses related work, Section 3 describes the construction of query templates, Section 4 reports the experiment design and protocol, the results of the experiments are analysed in Section 5, Section 6 provides a qualitative analysis of several user queries, finally Section 7 concludes the paper and outlines future research directions.

2. Related work

One of the central problems in the information retrieval process is the gap between a person's information need and a query statement that she submits to an IR system. It is usually difficult for people to express their information

JIS-0832-v2 Received: 24th September 2008 Revised: 9th December 2008

needs informally using a natural language, let alone to formulate them as effective queries for an IR system. Belkin et al. [2] 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." [2] (p. 62). Most interfaces to experimental and operational IR systems are developed based on the assumption that it is difficult or impossible for users to express their information needs precisely. Thus, for example, most commercial search engine interfaces have small query textboxes that invite short queries. On the other hand, experiments using test collections, such as TREC, showed that query length has a positive effect on performance [3]; therefore it is worth encouraging users to submit more detailed queries.

While in operational IR settings, the average query length is 2-3 words [4], a number of user studies showed that users are willing to submit longer and more expressive queries when encouraged to do so [5, 6]. In a user study by Belkin et al. [5], an experimental interface, which encouraged users to submit a longer query by providing them with a larger query textbox, was compared to a standard interface. The study proved that users submitted longer queries when using the experimental interface, and that reported user satisfaction was higher for the experimental interface than the control interface. However, no correlation between query length and performance was observed.

Kelly et al. [6] conducted an experiment within the framework of the HARD (High Accuracy Retrieval from Documents) track of TREC 2004. The HARD 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. In their clarification form Kelly et al. asked the user to supply information in response to the following questions/prompts: "Describe what you already know about the topic."; "Why do you want to know about this topic?" and "Please input any additional keywords that describe your topic." These prompts were followed by large textboxes, inviting long answers. The questions/prompts were intended to encourage the user to talk more about his/her information need. The results point to a relationship between query length and performance, and indicate that user-supplied responses improved retrieval performance compared to their initial queries. The systems in the above studies encourage and prompt users to talk more about their information needs, but they do little to help them articulate the relationships between concepts and entities that are important to their information needs.

Price et al. [7] proposed the semantic components model, which is intended to represent document contents in terms of various domain-specific topic aspects. Examples of semantic components in the medical domain are etiology and treatment. They developed and evaluated an experimental interface for a medical domain search system. Users were invited to enter their query plus fill in any textboxes representing various domain-specific semantic components. Documents in their collection were annotated with the semantic components, and an IR system ranked documents based on the match between the user-specified semantic components, and the semantic components assigned to documents. They reported improved performance compared to the control system, which did not use semantic components in query formulation and document ranking.

One of the contributions of our research is investigation of the template-based query formulation under domain-independent search conditions. The results of the user study suggest that while users found the process of template-based query formulation not as easy as using single-textbox interface, the queries they built tended to enhance the retrieval performance of a bag-of-words IR system.

The problem of satisfying complex information needs is relatively unexplored, but is beginning to gain attention. Most of the experimental research was conducted over the last few years within the framework of three TREC initiatives: the Relationship task of the Question Answering 2005 track and Complex Interactive Question Answering (CIQA) 2006 and 2007 tasks [8]. One of the aims of CIQA was also to study the interactive aspect of the IR process within the context of the complex question answering task. The opinion retrieval task in the Blog track [9] of TREC

JIS-0832-v2 Received: 24th September 2008 Revised: 9th December 2008

focused on one type of complex information needs – finding documents containing positive and negative opinions about sought entities.

Over the last few years, active research has been conducted in the area of automatic identification and classification of semantic relations. Various evaluation frameworks, such as SemEval [10] and Recognizing Textual Entailment (RTE) task [11] in the Text Analysis Conference (TAC), promote the development of new techniques, which may lead to the development of IR systems that make use of semantic relations between entities in document ranking. Therefore, it is important to investigate how users can communicate to an IR system the semantic relations between entities that they are interested in. Some recent work in the field of semantic relation identification and classification is summarized below.

Hearst [12] proposed the idea of mining a large corpus for lexical patterns that would help to identify a specific relation (hyponym-hypernym). Turney and Littman [13] classify noun-modifier pairs according to the semantic relations between them. They use counts of various short joining terms, obtained by web queries to measure the similarity between the test pair and the training pairs. Nakov and Hearst [14] propose a method that mines the Web to build a set of features that frequently occur in the same sentences as the target word pair. They then calculate similarity using Dice coefficient between the feature vector representing the test example with the vectors representing the training examples. Beamer et al. [15] used syntactic, lexical and semantic features from various knowledge bases. For example, they used information from WordNet and other sources to determine whether a noun belongs to spatial or temporal category. This served to recognize near-miss examples in cause-effect relationships, such as "activation by summer". Within the area of opinion detection, Hurst and Nigam [16] propose a method of identifying sentences that are relevant to some topic and express opinion on it. Yi et al. [17] propose to extract positive and negative opinions about specific features of a topic. By feature terms they mean terms that have either part-of or attribute-of relationships with the given topic or with a known feature of the topic. Yang et al. [18] relied on the opinion terms, rare non-dictionary terms, and phrases involving "I" or "you" pronouns compiled semi-automatically from a training dataset.

3. Constructing query templates

In order to identify a set of possible query templates, we have manually analyzed the descriptions and narratives of 100 TREC topics (#301-400) from TREC 6 and 7 ad-hoc tracks. The goal of this analysis was to identify a set of domain- and topic-independent semantic relations between concepts, about which users want to find information. We used the Rhetorical Structure Theory [19] to guide us in the analysis of the TREC topic narratives.

Rhetorical Structure Theory (RST) is a framework for the description and analysis of texts in terms of functions and roles played by different segments of text. Each natural language text has some coherence or unity, which means that every segment, such as a clause or sentence, has a certain role or function in text. Consider the following sentence: "Jack couldn't get on the plane. He lost his ticket." In this example, the first sentence represents an effect, while the second one represents a cause. Compare the previous example with the following: "Jack couldn't get on the plane. He likes apples." These two sentences lack coherence because it is not clear to the reader why they are placed together.

The RST analysis consists of identifying coherence relations between the segments of text, i.e. the role that one segment plays in relation to the other. In most cases, one of the segments is more essential to the overall communicative goal of the text, and is called *nucleus*, while the other, less essential, called *satellite*. In the first example above, the relation between segments is "Cause". Here, the effect (that Jack couldn't get on the plane) is

JIS-0832-v2 Received: 24th September 2008 Revised: 9th December 2008

more important than the cause (that he lost his ticket), therefore, the effect is the nucleus and the cause is satellite. Such RST relations are called nucleus-satellite relations. Another kind of relations is multinuclear, in which neither of the relations is more central to the communicative goal of the text. An example of such relation is "Contrast", whereby two nuclei are compared and contrasted.

The nucleus-satellite relations can be of two types: presentational and subject-matter. "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 that the reader recognizes the relation in question" [20]. An example of a presentational relation is "Evidence", which means that a reader's understanding of the satellite increases his/her belief/acceptance of the nucleus. Our focus was on identifying subject-matter relations between concepts in the text of TREC topic descriptions and narratives. The RST subject-matter relationships include: Circumstance, Condition, Elaboration, Evaluation, Interpretation, Means, Non-volitional Cause, Non-volitional Result, Otherwise, Purpose, Solutionhood, Unconditional, Unless, Volitional Cause, Volitional Result.

The analysis of the TREC topics was done by one of the authors of the paper. After reading the text of the description and narrative for a topic, the text was annotated with the RST subject-matter relations. The number of instances of each relation was recorded. We dropped four infrequent relations: "Circumstance", "Condition", "Unconditional" and "Unless", and added "Contrast" multinuclear relation. For each of these relations we then wrote one to two query templates in the form of fill-in-the-blanks questions. The phrasing of the templates was inspired by the actual questions that TREC topic authors used in the topic descriptions and narratives.

Table 1. Question templates and corresponding RST relations. $N_{completed}$ – the number of users who completed each template in our user study.

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 total, 13 query templates representing general semantic relations between concepts were written as a result of our analysis. The templates and the corresponding RST relations are presented in Table 1. As can be seen from the

JIS-0832-v2 Received: 24th September 2008 Revised: 9th December 2008

table, there is no one-to-one mapping between the RST relations and the developed templates. 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 ____?" 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 users are expected to interact with the templates by completing those they find relevant to their information needs.

We realise that the RST relations and the fill-in-the-blanks questions, into which they were cast, do not represent all possible entity relations about which users may want to find information. This set of relations, however, is broad enough to investigate the research questions posed by our study.

4. User study

4.1. Experiment Design

The goal of the user study is two-fold: (1) to investigate whether users are able to articulate complex information needs using query templates and (2) to determine how effective the template-based queries are when used with a bag-of-words IR system.

To investigate these questions, the following types of data were gathered in the course of the user study:

- Retrieval performance values based on the users' binary relevance judgments (Precision at 5, 10 and 15 documents retrieved in response to the users' queries);
- Search session logs (e.g., time to formulate the query, time to read/judge documents, number of query terms entered, number of templates filled in);
- The users' subjective comments (e.g., satisfaction with the search results, satisfaction with the overall system, familiarity with the topic prior to search) elicited via a questionnaire.

The data gathered via search session logs was used to measure the users' ability to formulate queries using templates, while the retrieval performance values and users' reported satisfaction levels were used to measure the retrieval effectiveness of template-based query formulation.

Two search interfaces were developed: the control and the experimental (template-based). The control interface contained one large textbox, where the user was invited to enter any words, phrases and sentences representing his/her information need (Figure 1). The textbox of the control interface was deliberately made large to encourage entry of long queries. The experimental interface consisted of 13 topic-independent query templates (Figure 2). The retrieval system used with both interfaces was Okapi with the default b and k1 values [21].

Thirty users, who are graduate students at the University of Waterloo, participated in the study. The user demographic details elicited through a short entry questionnaire are summarised in Table 2. As the search tasks, we used 10 topics from the TREC HARD 2004 track [22] collection. The topics are not the same that were used to create the templates using RST analysis. The first criterion for the selection of topics was that they represent a complex information need, which was taken to mean a case where the topic narrative contains at least one general semantic relation. The second criterion was that topics had 10 or more relevant documents in the HARD 2004 collection. This

JIS-0832-v2 Received: 24th September 2008 Revised: 9th December 2008

was done to eliminate topics with insufficient coverage in the collection. An example of the selected topic is given in Table 3. The collection on which the search tasks were performed was HARD 2004, consisting of 635,650 documents (newswire articles).

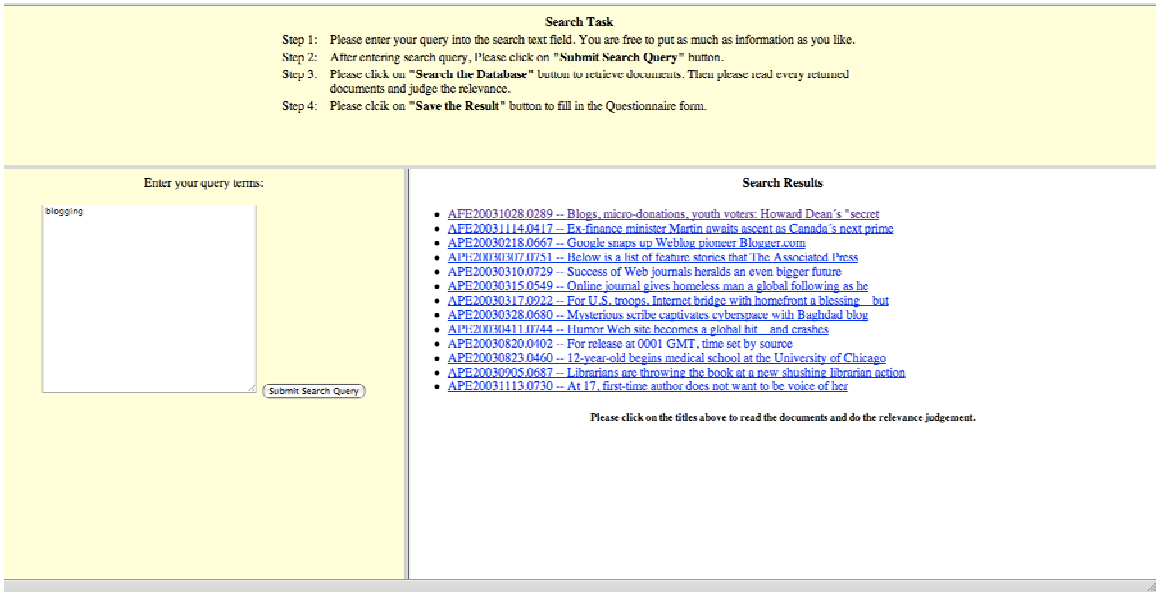

Figure 1. The control interface.

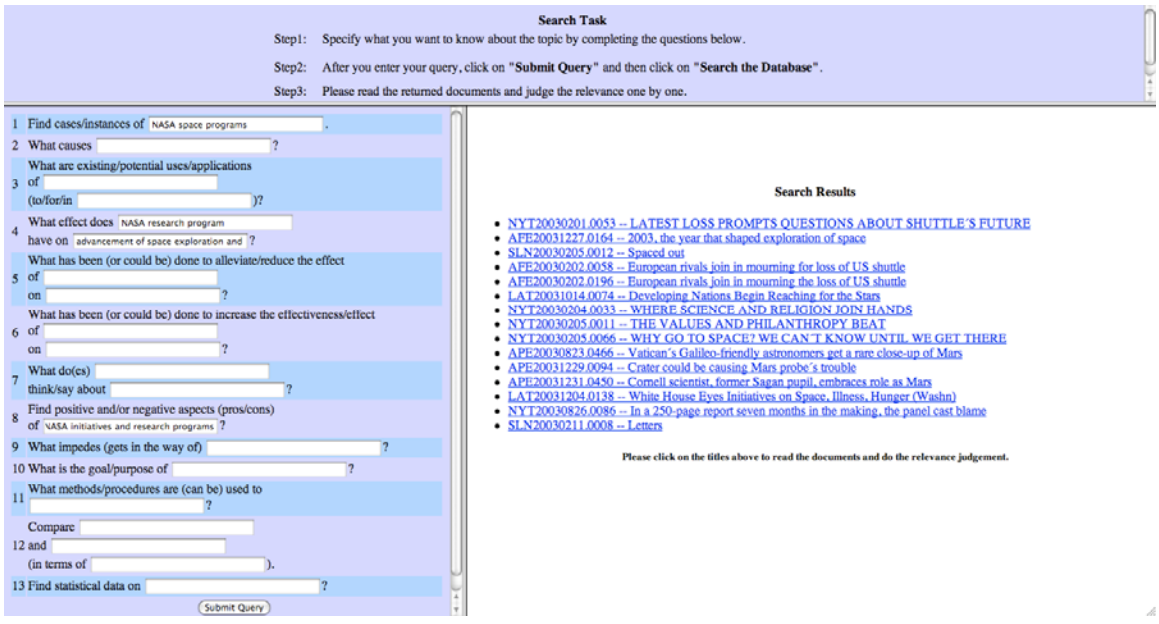

Figure 2. The experimental (template-based) interface.

JIS-0832-v2 Received: 24th September 2008 Revised: 9th December 2008

Table 2. 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

Table 3. Example of the selected TREC topic

Number: 406
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.

The experiment was designed using the Latin Square principle. 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. The rationale for using TREC topics as search tasks, as opposed to letting the users think of their own information needs, is that by giving the same topic to 3 users with the control system, and 3 – with the experimental, we are able to control the effect of the topic on the retrieval performance.

JIS-0832-v2 Received: 24th September 2008 Revised: 9th December 2008

Table 4. Experiment design

User #	Search task 1		Search task 2	
	Interface	Topic	Interface	Topic
1	Control	A	Experimental	B
2	Experimental	A	Control	B
3	Control	B	Experimental	A
4	Experimental	B	Control	A
5	Control	A	Experimental	B
6	Experimental	A	Control	B

4.2. Experimental Protocol

Upon arrival each participant read and signed the informed consent letter and was given a short entry questionnaire, followed by a brief tutorial of the first system (experimental or control). The participant was then shown the TREC title, description and narrative of the assigned topic. Then, he/she was presented with 3 relevant documents for this topic, which were randomly selected from HARD 2004 relevance judgments file. The relevant documents were shown to the participants in order to increase their level of familiarity with the topic, since we hypothesize that template-based interface would be helpful to expert users with complex information needs. Such users already have a high level of familiarity with the subject and may be interested in learning about specific relationships between concepts or entities in their domain of interest. On the contrary, somebody having little or no familiarity with the topic is unlikely to benefit from the use of query templates, as he may simply have insufficient knowledge to identify related concepts and formulate complex queries.

The user was then presented with the corresponding query formulation interface (experimental or control) and was asked to formulate a query based on the topic given to her. The text of the topic was not displayed on the screen at the query formulation stage (see Figures 1 and 2) to prevent copying and pasting, but was given to the user on paper. After the user submitted the query, the query terms were stemmed, and document retrieval was performed using Okapi BM25. The three relevant documents shown to the user earlier were excluded from the search results. We also excluded duplicate documents from the retrieved set. The top 15 retrieved documents were shown to the user, who then judged each of them as relevant or non-relevant. At the end of each search task the user was given a short questionnaire eliciting feedback regarding the system used for that task. After a short break the user proceeded to the second search task. The experimental protocol for the second search task was the same as for the first. The experiment received full ethics clearance from the Office of Research Ethics at the University of Waterloo.

5. Results and discussion

5.1. Retrieval Performance

The performance of the user queries formulated by means 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.

JIS-0832-v2 Received: 24th September 2008 Revised: 9th December 2008

Table 5. 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 can be seen from the table, the template-based interface helped users to formulate on average better queries that led to higher performance in all measures. Performance of the template-based system 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. The improvement in P@10 is statistically significant.

5.2. Query and Search Session Characteristics

We have logged different parameters of the user search sessions, namely: time spent reading the 3 sample relevant documents shown to users prior to each search session, time spent formulating the query, time spent 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. Table 6 summarizes the average values of these parameters. As we can see, the users spent on average 27 seconds more formulating the query using the templates, compared to entering the query into one textbox in the control interface. This observation is not surprising, since 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.

Table 6. Characteristics of the search sessions using control and template-based interfaces

	Control (mean±stdev)	Template-based (mean±stdev)
Time spent reading 3 sample relevant documents (min:sec)	09:28±06:10	08:18±05:31
Time spent formulating the query (min:sec)	04:08±03:39	04:35±03:36
Time spent reading and 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

An interesting parameter is the number of query terms entered. The users on average entered 16.17 query term tokens (including stopwords) using the control interface, and 31.83 using the template-based interface. The difference (96.91%) is statistically significant (t-test, p=0.0002). The number of query terms without duplicates and stopwords is 10.2 (control) and 12.03 (template-based). The difference (17.97%) is not statistically significant. Interestingly, on average 19.8 query terms entered using the experimental interface are either duplicates of the words previously entered 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. This is not entirely unexpected since the user may be interested in different aspects of the same concept and enter it into several templates. Nevertheless, 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. The average *idf* of query terms is slightly higher (4.4%) for the control interface, but the difference is not statistically significant.

JIS-0832-v2 Received: 24th September 2008 Revised: 9th December 2008