Cognitive Search Intents Hidden Behind Queries: A User Study on Query Formulations

Makoto P. Kato, Takehiro Yamamoto, Hiroaki Ohshima, and Katsumi Tanaka
Kyoto University, Yoshida Honmachi, Sakyo, Kyoto, Japan 6068501
{kato,tyamamot,ohshima,tanaka}@dl.kuis.kyoto-u.ac.jp

ABSTRACT

This study investigated query formulations by users with *Cognitive Search Intents* (CSI), which are needs for the cognitive characteristics of documents to be retrieved, *e.g.* comprehensibility, subjectivity, and concreteness. We proposed an example-based method of specifying search intents to observe unbiased query formulations. Our user study revealed that about half our subjects did not input any keywords representing CSIs, even though they were conscious of given CSIs.

Categories and Subject Descriptors

H.3.3. [**Information Search and Retrieval**]

Keywords

cognitive search intent; query formulation; Web search

1. INTRODUCTION

The estimation of search intents has been intensively tackled in recent years, and a better understanding of search intents has been considered as one of the most demanding challenges for modern search engines. Although search intents on the topic of documents (or *topical search intents*) have been dominantly addressed in the literature, not only the topic but also the *cognitive characteristics* of documents can be specified by users' search intents. The cognitive characteristics of documents are defined as the document characteristics perceived by readers, and these include comprehensibility, subjectivity, and concreteness. *Cognitive Search Intents* (CSIs), which are users' needs for the cognitive characteristics of documents, can be seen in a wide variety of searches, *e.g.* "I want to find *comprehensible* documents on black holes," "a *concrete* explanation of monopolies," or "documents *subjectively* written about Black Berry."

We investigated users' query formulations for CSIs by conducting a questionnaire-based user study. Our user study revealed that about half our subjects did not input any keywords representing CSIs, even though they were conscious of given CSIs. Our findings suggest users over-adapt to current Web search engines, and create opportunities to estimate CSIs with non-verbal user input.

WWW'14 Companion, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2577279.

2. METHODOLOGY

When we try to observe users' search behaviors, the most popular method is to present a task description and to ask subjects to conduct the task. However, this methodology has a serious drawback in that users' query formulations can be highly biased by the task description. For example, if we explain a task as "please find comprehensible documents on black holes," the user would be likely to input the query "comprehensible black holes" despite unfamiliarity with the term "comprehensible." Thus, we may not be able to observe natural, usual query formulations when the task is explained through text.

Therefore, we came up with an implicit, non-verbalized way of specifying CSIs that presents two types of examples to the subject and enables her/him to understand the CSIs. The first type of example is called a *positive* document set that consists of documents relevant to the CSIs we want to convey to the user, while the second type is called a *negative* document set that is composed of documents irrelevant to the CSIs. When we want the subject to search with the intent of "I want to find comprehensible documents on black holes," for example, we use comprehensible documents as the positive document set and incomprehensible ones as the negative document set. Having presented the two types of examples to the subject, we can ask him/her to search for documents on black holes that are not similar to the negative set but to the positive set.

We selected six types of CSIs and 40 topics (10 for exhaustiveness, 10 for comprehensibility, 10 for subjectivity and objectivity, and 10 for concreteness and abstractness), and assigned two topics to each subject in our user study, in which 1,800 subjects were recruited through an online-questionnaire company in Japan. The entire process for the questionnaire was completed on the Web, and all the instructions, questions, and topics were written in Japanese. We asked the questions below after instructing the search intents by the example-based method: **Q1**. How would you describe the given search intent to somebody else? **Q2**. What kinds of queries do you input to find documents relevant to the given search intents? **Q3**. (After showing a *sample query*[1]) How do you reformulate the shown query to find documents more relevant to the given search intents?

Note that different topics were used in **Q2** and **Q3**, while the CSIs used were the same. Thus, we assigned two types of search intents, and obtained a verbalized search intent for a search intent as well as two queries for both of the search intents.

[1]A sample query is a set of keywords that only represent topics of the given search intent.

Table 1: Query types for CSIs.

Cognitive search intent	#	Query type (%) direct	trans.	none	other
Exhaustive	134	0.7	36.6	**62.7**	0.0
Comprehensible	166	13.3	32.5	**53.0**	1.2
Objective	114	2.6	31.6	**65.8**	0.0
Subjective	128	0.8	**62.5**	36.7	0.0
Concrete	124	5.6	44.4	**49.2**	0.8
Abstract	84	0.0	45.2	**54.8**	0.0
Total	750	4.5	41.6	**53.5**	0.4

Three assessors labeled responses to filter out those who could not understand a search intent that we tried to inform them of. Responses were excluded if at least two assessors agreed. The inter-rater agreement was measured by using the average of Cohen's kappa coefficient between all pairs of assessors, and this was considered to be substantial ranging from 0.63 to 0.75. As a result, 375 (20.8% of the original responses) responses remained.

The assessors were also instructed to classify queries into the following types: *direct* (a term representing the CSI directly used in the query, *e.g.* "*concrete* monopoly"), *transformed* (a term representing a CSI somehow transformed in the query, *e.g.* "*example* monopoly" and "*review* blackberry"), and *none* (no term in the query related to the CSI, *e.g.* "blackberry"). The inter-rater agreement was measured in the same way as before: 0.72 for the query input (**Q2**) and 0.62 for the query reformulated (**Q3**). The query type was decided by votes, *i.e.* we used the query type labeled by two or more assessors. A few queries (0.4%) were labeled as *other*, since the three assessors labeled them differently.

3. FINDINGS

Table 1 lists the percentages for query types for CSIs. Note that we merged queries that were input for **Q2** and **Q3** in the questionnaire, as there were only small differences. Overall, the fraction of *direct* queries was quite small, while that of *transformed* and *none* queries were dominant in this categorization. When we compared these two dominant query types, there were more *none* queries than *transformed* queries. It follows that users were not likely to input a term that directly represented their CSIs, but were more likely to transform such a term into another term that they thought would be effective for retrieving documents relevant to their CSIs, or to only use keywords related to a topic that they wanted to read about that was not a CSI. Many *none* queries posed huge challenges in estimates of search intents regarding how *silent* CSIs could be detected.

Figure 1 compares the differences between verbalized search intents (**Q1**) and input queries (**Q2**) in terms of their component words. "Nouns (overlapping)" indicates terms that are included in both the verbalized search intent and query, while "Nouns (unique)" indicates terms that appear only in a verbalized search intent but not in its query, or *vice versa*. It is obvious that queries contain few verbs and adjectives. As the cognitive characteristics of documents are often verbalized in the form of adjectives, this trend might prevent users from explicitly inputting their CSIs. Moreover, it can be seen that about two nouns in the verbalized search intents were not used in subjects' queries, while on average 1.3 nouns in the queries were not used in their verbalized search intents. This finding might imply that users input some of the nouns from their verbalized search intents; on another front, they generate nouns suitable for keyword queries as alternatives to unused nouns, verbs, and adjectives. Although not conclusive, this hypothesis plus a large portion of *transformed* queries together sketch the shapes of

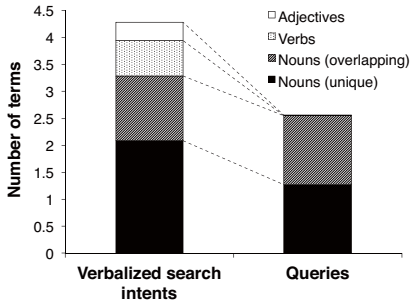

Figure 1: Difference between verbalized search intents and queries in terms of their component words.

query formulations for CSIs: users usually avoid inputting CSIs as adjectives, and translate their CSIs into nouns.

4. CONCLUSIONS

This study investigated users' formulations of queries for CSIs. Our user study revealed that about half our subjects did not input any keywords representing CSIs, even though they were conscious of given CSIs.

Our results demonstrated that half the subjects did not include terms representing CSIs in their queries, even though their verbalized search intent included such terms (this was ascertained by three assessors). A possible explanation for this phenomenon is that subjects over-adapted to the Web search engine. An early study on Web searches conducted by Pollock and Hockley found that some novices tried to enter natural language queries [3]. In addition, Bilal reported that 35 % of 22 seventh-grade children tried to search with a natural language question [1]. On the other hand, experienced users do not usually input natural language queries into search engines. Thus, the way users formulate queries might be acquired through experience with Web searches. Putting these findings all together, our hypothesis is that few experienced subjects input keywords related to CSIs because they knew Web search engines could not effectively process such words through their experience with search engines

The finding also creates opportunities to estimate CSIs with non-verbal user input. Although much work has utilized query logs to investigate users' search activities, the problem with silent CSIs suggests that only query-log-based analysis is not adequate for detecting users with CSIs. Therefore, clickthrough and interaction data are necessary to precisely estimate CSIs as can be seen in some previous work (*e.g.* [2]).

5. ACKNOWLEDGMENTS

This work was supported in part by the following projects: Grants-in-Aid for Scientific Research (Nos. 24240013 and 24680008) from MEXT of Japan, and Microsoft Research CORE Project.

6. REFERENCES

[1] D. Bilal. Children's use of the yahooligans! web search engine: I. cognitive, physical, and affective behaviors on fact-based search tasks. *Journal of the American society for information science*, 51(7):646–665, 2000.

[2] Z. Cheng, B. Gao, and T. Liu. Actively Predicting Diverse Search Intent from User Browsing Behaviors. *Proc. of WWW*, pages 221–230, 2010.

[3] A. Pollock and A. Hockley. What's wrong with internet searching. In *D-Lib Magazine*, 1997.