Questions vs. Queries in Informational Search Tasks

Ryen W. White, Matthew Richardson, and Wen-tau Yih
Microsoft Research
Redmond, WA 98052 USA
{ryenw,mattri,scottyih}@microsoft.com

ABSTRACT

Search systems traditionally require searchers to formulate information needs as keywords rather than in a more natural form, such as questions. Recent studies have found that Web search engines are observing an increase in the fraction of queries phrased as natural language. As part of building better search engines, it is important to understand the nature and prevalence of these intentions, and the impact of this increase on search engine performance. In this work, we show that while 10.3% of queries issued to a search engine have direct question intent, only 3.2% of them are formulated as natural language questions. We investigate whether search engines perform better when search intent is stated as queries or questions, and we find that they perform equally well to both.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *search process*; *selection process*.

Keywords

Query formulation; Natural language queries; Informational search

1. INTRODUCTION

Web search engines have been optimized to handle keyword-based queries. However, recent studies have identified an increase in the fraction of queries phrased as natural language questions, e.g., [*why is the sky blue?*][7]. Search engines have been designed to handle short keyword queries, e.g., [*blue sky reason*], so the apparent evolution in how searchers express their information needs warrants deeper study. Important questions include: how prevalent are questions and question-answering intentions in Web search? When people have intentions expressible as questions, what benefit comes from formulating them as keyword queries or questions directly?

The query formulation process has been studied in detail in information retrieval [1][2][4]. Keyword queries can be challenging to formulate in some situations, especially when information need are vague or the searcher is inexperienced [4]. Researchers have explored ways to encourage people to provide richer queries by requesting longer query statements [2], as well as investigated costs associated with query reformulation and result examination [1]. However, this research has not focused on the implications (in terms of result relevance) of different query formulation strategies, an important decision that searchers must make for every query. Question-answering has also been compared against information retrieval (IR) methods [6], but using specialized question answering systems and not generic Web search engines as we study here.

In this paper, we study query formulation strategies in Web search, with an emphasis on different formulations of the same informational intent, expressed as keyword queries (*query*) or natural language questions (*question*). We make the following contributions:

WWW 2015 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2742769

- Measure the occurrence of question-answering intent, including quantifying the prevalence of expressing this intent as natural language question queries and as keyword queries, and;
- Compare the quality of search engine results when formulating an informational search intent as a query vs. a question.

2. QUESTION PREVALENCE

To better understand the importance of answering questions submitted to search engines, we computed their prevalence in the logs of a commercial search engine. For details on the method for determining whether a query is a question, see our technical report [9]. We considered English queries from November 2011 through January 2013 that originated from the United States and ignoring queries automatically identified as spam or bot-generated (this is also the query set used in the remainder of this paper). Our findings are: 3.18% of the queries were written in natural language question form. This is greater than the 1.8% found in an earlier study [7], with differences likely related to the timeframe and the exact question definition. Also, corroborating [7], we found 2.34% of the queries were expressed as questions in the timeframe of May 2010 through July 2011. While the average length of keyword queries with a question-answering intent is 3.80 words, questions averaged 7.39 words. Thus, searchers are investing considerable additional time in generating question-based query statements.

3. SEARCH INTENT

To understand the nature of search intents observed in our logs, and to create a set of informational search tasks for further analysis, we performed labeling of the search intentions. We adopted the query classification taxonomy of Rose and Levinson [8] with some minor modifications. At the top level, queries are categorized into *navigational*, *informational*, or *resource*. Queries with question-answering intent appear in the informational category, which is further divided into *directed*, *undirected*, and *other*. The directed category refers to any query where the searcher seeks to learn something particular about a topic, as compared to the undirected category where they seek to learn about a topic in general. Finally, the directed category is split into *closed* vs. *open*, indicating whether the question can be answered with a single unambiguous answer (closed) vs. is more open-ended. For instance, [*how many calories are in a cup of flour*] is informational-directed-closed whereas [*why are calories bad for you*] is informational-directed-open. The resource category is divided into *virtual* or *physical*, indicating the type of resource being sought. We added four more "junk" categories of *pornography*, *other*, *cannot tell*, and *error or non-English*.

3.1 Task Judging and Results

We randomly sampled 1000 search sessions from the query log and asked judges to categorize the first query into our task hierarchy, based on the initial query, subsequent refinement queries in the session, and associated clicks on search results (as in [8]). Often, the intent of in initial query was unclear without the context of the subsequent session. We employed crowdsourced judges from Clickworker.com, provided under contract. Judges resided in the U.S. and were required to be fluent in English. Each query was evaluated by 10 judges and inter-judge agreement as measured by Fleiss' kappa was 0.357 (considered to be *fair* agreement). The final label

Table 1. Percentage of judged queries in each query category.

Query Category	*% Queries*
Navigational	54.4
Informational	31.8
Directed	10.3
Closed	5.3
Open	5.0
Undirected	14.3
Other	7.2
Resource	6.9
Virtual	4.8
Physical	2.1
Pornography	2.7
Error/Other	4.2

for a query is the mode of the labels, with ties broken randomly. The frequency of each category is given in Table 1. Note that while 3.2% of the queries are easily identifiable as questions based on the initial word or terminal question mark, 10.3% of queries were labeled as having a question answering (informational-directed) intent. That is, in approximately 70% of occasions a question-answering intent arrives at the search engine, it is formulated as a keyword query. For a detailed comparison of the differences between our findings and those of Broder [3] or Rose and Levinson [8], please see our technical report [9].

4. RESULT RELEVANCE

Given the significant portion of queries with informational-directed intents, we sought to understand whether formulating the intent as keyword queries led to better result relevance than natural language questions. We ran two crowdsourcing tasks to answer this question.

4.1 Crowdsourcing Tasks

In the first task (*query formulation*), workers were given a specific information need, or *intent*, as a search task statement. They were then asked to compose a keyword- or question-based query statement. To measure the effectiveness of these formulations we issued them to two popular commercial Web search engines, denoted A and B, using their provided public APIs. The second task (*relevance judgment*) involved judging whether the search results were relevant to the original intent, defined by the search task statement.

For each of the 103 tasks identified with the *informational-directed* search intent, the authors created a search task description after examining the session queries. For instance, a session that starts with the query [*rule of standard form*] becomes the search task: "*You are reviewing some linear algebra materials and encounter the rule of standard form. Find out its meaning.*" When creating such tasks, each statement typically consists of two sentences: one providing general background scenario on why such an information need may arise, and one further specifying the exact required information.

Based on these search tasks, three query formulation crowdsourcing tasks were developed. These share the same interface, except in the description where we requested different query types: *keyword-based*, *question*, and *question for search engines*. By specifically stating that the questions will be used as search engine queries, we sought to understand whether this affects question construction. Each crowdsourcing task was assigned to 10 different workers. Moreover, to ensure that the search task is new to the worker each time, workers could not see the same search task description more than once when working on different types of query formulation.

After collecting the queries/questions formulated by workers, we issued them to both Web search engines, and retained the top three results. Another set of five workers assessed the relevance of each result to the original *task* description on a five-point scale: *perfect*, *excellent*, *good*, *fair*, and *bad*. The query/question used to obtain the results was hidden; only the task description was shared.

4.2 Relevance Results

Prior to analysis, we removed seemingly erroneous queries and relevance judgment labels. We employed several methods to identify careless workers, such as by examining queries or questions they entered, and by comparing their task time with the average. This data cleaning removed 30% of queries and questions. In addition, when determining the final relevance judgment of each pair of task and Web page, we use the mode as the final judgment. For ties, we used the average of the multiple modal values. Given the judgments, we computed the normalized discounted cumulative gain (NDCG) [5] for the two systems, and report the average in Table 2.

Table 2. Relevance results (in terms of NDCG) of keyword queries and natural language questions.

Engine	*Query*	*Question$_{Engine}$*	*Question$_{Any}$*
A	0.471	0.465	0.462
B	0.493	0.487	0.497

Our analysis of different configurations shows that when formulating an informational-directed search intent directly as a natural language question, result relevance (in NDCG) is statistically indistinguishable from that of traditional keyword queries (analysis of variance, $F(5,611) = 0.520$, $p = 0.762$). Table 2 shows that we observe the same phenomenon on both engines. The similarity of the NDCG values for *Question$_{Engine}$* and *Question$_{Any}$* suggests that considering the intended target of the question during query formulation (target =search engine vs. target=anywhere) has little impact on relevance.

5. DISCUSSION AND CONCLUSIONS

Question queries are common, but most (around 70%) of informational-directed intentions are represented in keywords. We study whether these intentions are better formulated as questions or as queries, and found little difference in relevance. Since questions offer little benefit, searchers should only be utilizing keyword queries. However, human behavior is not strictly rational, and the prevalence of natural language queries continues to increase. There may be other driving factors, such as a desire to find answers on community question answer sites and interfaces that encourage the more natural expression of information needs, e.g., spoken dialog.

There are some key areas of future work. Rather than recommending to searchers that they adopt a particular strategy based on average search performance, mechanisms could predict the best strategy on a per query basis. This could form part of search support to engage searchers to elicit a natural language query should a question be predicted to perform better, suggest variants to searchers, or use the variants to perform query alterations, to improve relevance.

REFERENCES

[1] Azzopardi, L., Kelly, D., and Brennan, K. (2013). How query cost affects search behavior. *SIGIR*, 23–32.

[2] Belkin, N.J. et al. (2003). Query length in interactive information retrieval. *SIGIR*, 205–212.

[3] Broder, A. (2002). A taxonomy of web search. *SIGIR Forum*, 36(2): 3–10.

[4] Furnas, G.W. et al. (1987). The vocabulary problem in human-system communication. *CACM*, 30(11): 964–971.

[5] Järvelin, K. and Kekäläinen, J. (2002). Cumulative gain-based evaluation of IR techniques. *TOIS*, 20(4): 422–446.

[6] Laurent, D., Séguéla, P., and Nègre, S. (2006). QA better than IR? *EACL Workshop on Multilingual QA*, 1–8.

[7] Pang, B. and Kumar, R. (2011). Search in the lost sense of "query": Question formulation in Web search queries and its temporal changes. *ACL*, 135–140.

[8] Rose, D.E. and Levinson, D. (2004). Understanding user goals in Web search. *WWW*, 13–19.

[9] White, R.W., Richardson, M., and Yih, W. (2014). *Questions vs. Queries in Informational Search Tasks*. Microsoft Research Technical Report MSR-TR-2014-96, July 2014.