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 have traditionally required searchers to formulate information needs as a set of keywords rather than in a more natural form, such as questions. Recent studies have found that search engines are observing an increase in the fraction of Web search queries that take the form of 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 and searcher efficiency. In this work, we study the behaviors of search engines when handling keyword-based queries and natural language questions, as well as the costs incurred by searchers in creating query statements of each form. We show that although informational search intentions are often expressed as keyword queries, when given the same search intent expressed as a query and as a natural language question, search engines in fact perform equally well in terms of relevance. Since creating queries has been assumed to be challenging, this equality should support an increase in question-querying. However, question formulation has an associated cost, e.g., we show that generating natural language questions for search engines takes much longer than keyword queries for the same intent. Our findings suggest that searchers should stick with keyword queries and that the increase in question prevalence is related to factors beyond search engine performance.

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 query statements. However, recent studies have found that the fraction of queries submitted to search engines that take the form of natural language questions, e.g., [*why is the sky blue?*] is increasing [12]. 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 further investigation. Important questions include: how prevalent are questions and question-answering intentions in Web search? When people have intentions expressible as questions, what are the costs and benefits of formulating them as keyword queries or questions directly?

The query formulation process has been studied in detail in information retrieval [15]. Keyword queries can be challenging to formulate in some situations, especially when the information need is vague or the searcher is inexperienced [7]. Taxonomies of search intentions have been developed, including the well-known tripartite classification of navigational, informational, and transactional intents [5], and refined variants thereof [13], but searcher needs may have evolved in the time since those studies. Researchers have explored ways to encourage people to provide richer queries either by requesting longer query statements [4] or by asking searchers to contextualize their information needs along different dimensions, including topic familiarity [9]. Researchers have also investigated costs in search interaction compared with information gained over time [14], and the costs associated with query reformulation and result examination [3]. However, this research has not focused on the costs and benefits 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 [10], 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 a particular emphasis on different formulations of the same informational intentions. We seek to understand the benefits and costs from different query formulations of keyword queries versus natural language questions. We characterize the nature of search intents expressed in online searching, and show that such informational tasks occur frequently in search engines. Given this prevalence, crowdsourced judges were given a task and generated variants of the same intent: keyword queries (denoted *Query*) and natural language questions (*Question*). We estimated the benefits and costs of each formulation strategy. To understand the impact of the intended question audience, judges created two variants of *Question*: one for a search engine and one for general purpose (for anyone). We include these two variants in our analysis.

We make the following contributions with this research:

- 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 benefits and the costs connected with formulating queries and questions. We quantify benefits in terms of relevance and costs in terms of formulation and typing time.

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 (i.e., the fraction of search engine queries that are full questions). Using these logs also allowed us to mine example queries, useful for labeling search intents. To find these questions, we adopted the following approach: (i) Using the query log of a major search engine, we looked at all queries ending with a question mark; (ii) We then manually inspected the most commonly occurring initial terms for such queries, and added them to a dictionary of question indicators. Besides the 5W1H question types ("who", "what", "how", etc.) this list also included terms such as "does" and "should", and; (iii) Any query beginning with a question indicator or ending with a question mark was considered to be a question query. Inspecting the most frequent question queries revealed that two-word queries beginning with "will" were typically names (e.g., Will Smith). Hence, such queries were reclassified as keyword queries. Manual review of a sample of the resulting questions indicated a high accuracy of this technique for finding questions. Most of these question queries appeared to have informational intent and resulted in visits to Websites such as ehow.com, answers.yahoo.com, wikipedia.org, and answers.com. Some statistics of the question queries that appeared in our logs are provided in Table 1. The statistics were computed on English queries originating from the United States from two different time periods: May 2010 through July 2011 (*QL2011*), and November 2011 through

Table 1. Statistics on question prevalence in our query log. Differences from QL2011 to QL2012 are all statistically significant (p<0.001, using *Z*-tests of proportions for the first two rows and Mann-Whitney test for the last).

Statistic	*QL2011*	*QL2012*
Portion of queries that are questions	2.34%	3.18%
Portion of questions that end with "?"	15.6%	16.1%
Average words per question	7.18	7.39

January 2013 (*QL2012*). We ignored queries automatically identified as spam or bot-generated traffic.

The fraction of queries labeled as questions resembles earlier studies, e.g., 1.8% [12], with differences likely related to the timeframe and the exact question definition. Also, corroborating [12], the portion of queries that are questions appears to be increasing over time. It is interesting that many questions (over 15%) end with question marks, even though punctuation is ignored by the search engine.

What could be more costly to searchers is the length of question queries. From our later analysis, we estimate that the average length of keyword queries with a question-answering intent is 3.8 words. However, from Table 1 we can see that the average question length exceeds seven terms. Searchers are investing considerable time in generating question-based query statements. We explore the value of this investment in terms of result relevance later in the paper.

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 in the logs. A searcher may have underlying question-answering intent when issuing queries, even though it is expressed in keyword form. Among other things, our analysis allows to us to quantify how often this occurs.

We adopted the query classification taxonomy of Rose and Levinson [13] 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 user is seeking 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*. The differences between our taxonomy and [13] are based on preliminary experiments indicating that judges had difficulty differentiating some subcategories, and also based on our interest in questions (so *informational-list* is merged into *informational-directed-closed* since it also seeks an answer to a question).

3.1 Task Judging and Results

We randomly sampled 1000 search sessions from *QL2012* (the remainder of the paper considers only this data set) 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 [13]). Often, the intent of a query is unclear without the context of the session that follows it. 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 assigned to a query is the label that had the most judges (i.e., the mode of the set of judgments), with ties broken randomly. We experimented with more advanced methods such as [16], but obtained similar results; we therefore used the mode for simplicity. The frequency of each category is given in Table 2. One item of note is that while 3.2% of the queries are easily identifiable as questions based on the initial word or ending with question mark (see previous section), 10.3% of the queries were labeled as having question answering (informational-directed) intent. Put another way, approximately 70% of the time a question-answering intent arrives at the search engine, it will have been formulated as a keyword query.

Table 2. Percentage of judged queries in each query category.

Query Category	*% of 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/Indeterminate	4.2

Also surprising is the difference between these findings and previous work. Broder [5] found that navigational queries constituted 20-24.5% of query traffic, and this was refined to 11-15% by Rose and Levinson [13]. However, we find over half (54.4%) of the query volume has navigational intent. Our judges were given the same description for what constitutes a navigational query as those previous studies. One difference between our study and [5] is that our queries are session-initial queries from a random sampling of sessions, such as was done in [13], as opposed to a random sampling from all queries, as in [5]. As is suggested in [13], this may lead to an increased percentage of navigational queries since navigational sessions tend to be shorter. Given that the previous studies were conducted over 10 years ago, there may be a number of other likely explanations for this apparent change in web search behavior, including the growth of the Web (i.e., more distinct Websites may result in more navigational search queries), emergence of highly popular Websites that are popular navigational queries today (e.g., Facebook, YouTube, Gmail, none of which existed when the previous studies were conducted), or a change in how people access Web content, favoring searching over browsing today.

4. RESULT RELEVANCE

Given that we found that a significant portion of queries had informational-directed intents, it is important to understand how modern search engines perform for this type of information need. In particular, we would like to understand whether formulating the intent using keyword queries leads to better result relevance than directly issuing natural language questions. To answer this question, we designed two crowdsourcing tasks, described in the next section.

4.1 Crowdsourcing Tasks

The two tasks we designed for comparing keyword-based queries and natural language questions were *query formulation* and *relevance judgment*. In the first task, workers are given a specific information need, or *intent*, as a search task statement. They are then

asked to compose a keyword- or question-based query statement. To measure the effectiveness of these two formulations we issue them to two popular commercial Web search engines, denoted A and B, using their provided APIs. The second crowdsourcing task involved judging whether the search results were relevant to the original intent, defined by the search task statement.

We employed this crowdsourcing design because we needed a way to determine whether questions or keyword queries were handled better by search engines, but also required a source of queries and questions that are trying to answer the same task, so as to make the comparison unbiased. Studying success levels for queries and questions from sources such as search logs would not support this direct comparison, since searchers may use different formulation strategies depending on the nature (e.g., difficulty) of the search task.

4.1.1 Query Formulation

For each of the 103 tasks that have been identified with the *informational-directed* search intent, the authors created a search task description after examining the queries in the session. For instance, a session that starts with the query "*rule of standard form*" becomes the following 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. The first provides a general background scenario on why such an information need may arise. The second sentence then further specifies 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*. The last two are essentially the same as we anticipate that they both should be normal, natural language questions. Nevertheless, by specifically stating that the questions will be used as input to a search engine, we sought to understand whether this affects question construction. In the interface, the task description is initially hidden from the worker until they click a "show task" button, and another button needs to be clicked thereafter before they can start typing their query statement. This staged-design helps us to record the time each worker spends on comprehending the search task, as well as the time to formulate the query, and then enter it. In addition to the query/question, the worker also needs to provide feedback at the end of each task on the difficulty of formulating the query statement on a five-point scale, from "very easy" to "very difficult." Each crowdsourcing task is assigned to 10 different workers. Moreover, in order to ensure that the search task is new to the worker each time, workers were not able to see the same search task description more than once when working on different types of query formulation.

4.1.2 Relevance Judgment

After we collected the queries/questions formulated by workers, we issued them as search queries to both of the search engines used in our study, and retained the top three results. Another set of workers assessed the relevance of each result to the original *task* description on a five-point scale: *perfect*, *excellent*, *good*, *fair*, and *bad*. The actual query/question used to obtain the search results are hidden from the judges; only the task description is shared. Each task-URL pair was labeled by five crowdworkers.

4.2 Relevance Results

Given the nature of crowdsourcing, low-quality queries and questions may be provided by careless workers and judgments of some Web search results could be inaccurate. Prior to analysis, we removed seemingly erroneous query statements 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. After this data cleaning process, we examined the remaining data and found that on average approximately 30% of the queries and questions were removed by the cleaning. In addition, when determining the final relevance judgment of each pair of task and Web page, we use the mode (i.e., the label that appears most often) as the final judgment (we judged a sample of task-URL pairs and found that the highest correlation between our judgments and the crowdsourced judgments was achieved via the mode, vs. average or median). When there was a tie, we used the average of the multiple modal values instead. Given the judgments, we computed the normalized discounted cumulative gain (NDCG) [8] for the two systems for this query set, and reported the mean average values in Table 3.

Table 3. The relevance results of keyword-based queries and natural language questions in NDCG. For a given engine, the differences between the three types of queries are small, indicating results are insensitive to whether a user submits a query statement using keywords or a question. Differences are not significant using a two-way ANOVA.

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

Although different configurations lead to minor differences in the relevance of the results, these differences were not statistically significant using a two-way analysis of variance (ANOVA), with engine and query formulation method as factors ($F(5,611) = 0.520$, $p = 0.762$), i.e., when formulating an informational-directed search intent directly as a natural language question, the relevance of the search engine results are statistically indistinguishable from that of traditional keyword queries. As can be seen in Table 3, we observe the same phenomenon on both of the search engines studied. Interestingly, explicitly requiring that the questions be formulated for submission to search engines (*Question$_{Engine}$*) does not appear to affect the relevance of the results returned, compared with not imposing this restriction (*Question$_{Any}$*). Note the differences between these query formulations were extremely minor, at most limited to the inclusion of an additional stopword term in *Question$_{Any}$*.

5. QUERY GENERATION COSTS

Although we find that the relevance of queries and questions is similar, we wanted to understand the *costs* associated with formulating each type of query statement. We used the queries and questions from the process described in the previous section and computed:

1. Formulation time: We defined the formulation time as the time between the task being understood by the worker and then starting to type the query statement.

2. Perceived formulation difficulty: Judges were asked to rate the level of difficulty in creating the query statement on a five-point scale from easy to difficult, where higher meant more difficult.

3. Query length: The number of characters in the query statement.

4. Typing time: The time to enter the query statement in the interface (i.e., the time from the first to the last keystrokes).

Table 4 displays the value for each of these four metrics for query and question, including the variation in the question target (engine or general purpose, denoted *Any*). In addition to each metric, we also report the median overall query creation time (formulation time plus typing time). Since we were dealing with ordinal or non-normally distributed data we used non-parametric testing. In situations where the medians are similar, we also report the mean average for reader reference. The overall median task comprehension

Table 4. Median costs incurred in query creation. Mean average in brackets [ ] for difficulty and query length. Post-hoc significance vs. Query: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Metric	*Query*	*Question$_{Engine}$*	*Question$_{Any}$*
Perceived difficulty	2 [1.72]	2 [1.78]	1 [1.54]
Query length (chars)	25 [27.33]	43 [44.77]***	44 [45.39]***
Formulation time (secs)	3.62	4.98**	3.03
Typing time (secs)	9.71	15.48*	13.06*
Total time (secs)	14.61	21.66*	16.51

time was 3.18s. We do not include the task comprehension time in Table 4, since we focus on durations once the task was understood, and the times were highly similar between the three query variants.

Table 4 shows that although formulating a question without considering the engine target appears to be easier, the differences are not significant (Kruskal-Wallis test, $\chi^2(2) = 3.62$, $p = 0.164$). Questions are much longer than queries ($\chi^2(2) = 102.96$, $p < 0.001$), and take longer to type ($\chi^2(2) = 7.07$, $p = 0.029$, Dunn's post-hoc tests: both $p < 0.05$). Interestingly, if we also consider the time required to formulate the query statements, we see that formulating a question for a search engine takes the longest time (4.98s vs. 3.62s for a keyword query and 3.03s generally, both $p \leq 0.005$). One explanation for this is that there is additional overhead involved in selecting terms that are both present in Web content and discriminatory. This is reflected in the slightly higher difficulty value (rating = 1.78) for *Question$_{Engine}$* in Table 4. The last row shows that it takes around 7s longer to create questions but the benefits are not borne out in relevance (Table 3). Given that keyword queries perform as well as questions, but questions take longer to create, searchers may be more efficient if they stick with keyword queries.

6. DISCUSSION AND CONCLUSIONS

We showed that question queries are common, but most (around 70%) of informational-directed intentions are represented as keyword queries. We study whether these are better formulated as questions or as queries, and found little difference in relevance, but an increase in the costs associated with formulating natural language questions for search engines. Based on these findings, since questions have a higher cost, are not easier to create, and offer no more benefit, it seems reasonable to assert that searchers should only be utilizing keyword queries. However, human behavior is not strictly rational [1], and the prevalence of natural language queries continues to increase. There may be other factors that drive this activity. These include a desire to find answers on community question answer sites, misconceptions about how search engines operate (or put more positively, a lack of functional fixedness [6] about how search engines process queries) especially among search novices, and interfaces such as those supporting spoken dialog, that encourage searchers to express their information needs more naturally.

There are some limitations of this work. We targeted a small number of informational-directed search tasks from our intent labeling. A broader analysis is needed to determine the generalizability of our findings. In addition, in analyzing the data we observed that the formulations from crowdworkers were around 20-25% longer than their counterparts in the query logs (+1 term in keyword queries, +1.5 terms in questions). One explanation is that our query formulation task requested one query to capture the intent, when natural search scenarios typically involve query reformulation. Although the percentage gain is similar for both strategies, and we focus on their *relative* performance, further work is needed, including refining the crowdsourcing tasks to permit query reformulations, or conducting studies of informational directed search behavior in a more natural setting. There have been studies on query dynamics over the course of search sessions [2][10]. There are interesting opportunities in understanding how queries transition to and from questions, and the underlying motivations. In addition, without knowledge of search engine internals we cannot ascribe our findings to particular aspects of engine operation. We can observe the output of the engine and by running controlled experiments we can better understand their algorithms and perhaps more fully explain some of our findings. For example, the small differences in retrieval could be attributed to stopword handling; if engines aggressively strip them then questions will resemble queries, and generate similar results.

There are some additional areas of future work. For example, our relevance results reported are averages across all queries. There are some tasks where questions perform better, and some where queries perform better. Rather than recommending that searchers should adopt a particular strategy based on average search performance (e.g., always use keyword queries for informational tasks), mechanisms to predict the best strategy on an individual query basis may be useful. This could form part of search support to engage users to elicit a natural language representation for a keyword query should a question be predicted to be perform better for the current task, suggest variants to searchers, or use the variants to perform backend query alterations, with the goal of improving result relevance.

REFERENCES

[1] Ariely, D. (2008). *Predictably Irrational: The Hidden Forces that Shape Our Decisions*. Harper Collins

[2] Aula, A., Khan, R.M., and Guan, Z. (2010). How does search behavior change as search becomes more difficult? *SIGCHI*, 35–44.

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

[4] Belkin, N. J., Cool, C., Kelly, D., Lee, H.-J., Muresan, G., Tang, M.C., and Yuan, X.J. (2003). Query length in interactive information retrieval. *SIGIR*, 205–212.

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

[6] Duncker, K. (1945). On problem solving. *Psychological Monographs*. 58(5): 270.

[7] Furnas, G.W., Landauer, T.K., Gomez, L.M. and Dumais, S.T. (1987). The vocabulary problem in human-system communication. *CACM*, 30(11): 964–971.

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

[9] Kelly, D., Dollu, V.D., and Fu, X. (2005). The loquacious user: A document-independent source of terms for query expansion. *SIGIR*, 457–464.

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

[11] Nordlie, R. (1999). "User revealment"—a comparison of initial queries and ensuing question development in online searching and human reference interactions. *SIGIR*, 11–18.

[12] 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.

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

[14] Russell, D.M., Stefik, M.J., Pirolli, P., and Card, S.K. (1993). The cost structure of sensemaking. *SIGCHI*, 269–276.

[15] Taylor, R.S. (1968). Question-negotiation and information seeking in libraries. *College and Res. Lib.* 29: 178–194.

[16] Zhou, D., Platt, J., Basu, S., and Mao, Y. (2012). Learning from the Wisdom of the crowds by minimax entropy. *NIPS*, 2204–2212.