Search in the Lost Sense of "Query": Question Formulation in Web Search Queries and its Temporal Changes

Bo Pang **Ravi Kumar**
Yahoo! Research
701 First Ave
Sunnyvale, CA 94089
{bopang,ravikumar}@yahoo-inc.com

Abstract

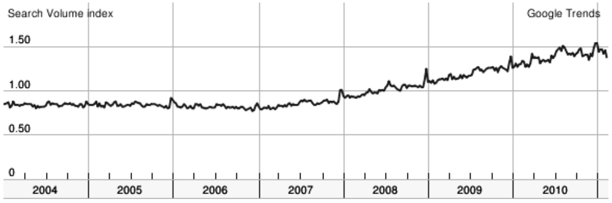
Web search is an information-seeking activity. Often times, this amounts to a user seeking answers to a question. However, queries, which encode user's information need, are typically not expressed as full-length natural language sentences — in particular, as questions. Rather, they consist of one or more text fragments. As humans become more search-engine-savvy, do natural-language questions still have a role to play in web search? Through a systematic, large-scale study, we find to our surprise that as time goes by, web users are more likely to use questions to express their search intent.

1 Introduction

A web *search query* is the text users enter into the search box of a search engine to describe their information need. By dictionary definition, a "*query*" is a question. Indeed, a natural way to seek information is to pose questions in a natural-language form ("*how many calories in a banana*"). Present day web search queries, however, have largely lost the original semantics of the word query: they tend to be fragmented phrases ("*banana calories*") instead of questions. This could be a result of users learning to express their information need in search-engine-friendly forms: shorter queries fetch more results and content words determine relevance.

We ask a simple question: as users become more familiar with the nuances of web search, are *question-queries* — natural-language questions posed as queries — gradually disappearing from the search vernacular? If true, then the need for search engines to understand question-queries is moot.

Anecdotal evidence from Google trends suggests it could be the opposite. For specific phrases, one can observe how the fraction of query traffic containing the phrase[1] changes over time. For instance, as shown next, the fraction of query traffic containing "*how to*" has in fact been going up since 2007.



However, such anecdotal evidence cannot fully support claims about general behavior in query formulation. In particular, this upward trend could be due to changes in the kind of information users are now seeking from the Web, e.g., as a result of growing popularity of Q&A sites or as people entrust search engines with more complex information needs; supporting the latter, in a very recent study, Aula et al. (2010) noted that users tend to formulate more question-queries when faced with difficult search tasks. We, on the other hand, are interested in a more subtle trend: for content that could easily be reached via non-question-queries, are people more likely to use question-queries over time?

We perform a systematic study of question-queries in web search. We find that question-queries account for $\sim 2\%$ of all the query traffic and $\sim 6\%$ of all unique queries. Even when averaged over intents, the fraction of question-queries to reach the

[1] www.google.com/intl/en/trends/about.html

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:shortpapers, pages 135–140,
Portland, Oregon, June 19-24, 2011. ©2011 Association for Computational Linguistics

same content is growing over the course of one year. The growth is measured but statistically significant.

The study of long-term temporal behavior of question-queries, we believe, is novel. Previous work has explored building question-answering systems using web knowledge and Wikipedia (see Dumais et al. (2002) and the references therein). Our findings call for a greater synergy between QA and IR in the web search context and an improved understanding of question-queries by search engines.

2 Related work

There has been some work on studying and exploiting linguistic structure in web queries. Spink and Ozmultu (2002) investigate the difference in user behavior between a search engine that encouraged questions and one that did not; they did not explore intent aspects. Barr et al. (2008) analyze the occurrence of POS tags in queries.

Query log analysis is an active research area. While we also analyze queries, our goal is very different: we are interested in certain linguistic aspects of queries, which are usually secondary in log analysis. For a comprehensive survey on this topic, see the monograph of Silvestri (2010). There has been some work on short-term (hourly) temporal analysis of query logs, e.g., Beitzel et al. (2004) and on long queries, e.g., Bendersky and Croft (2009).

Using co-clicking to infer query-query relationships was proposed by Baeza-Yates and Tiberi (2007). Their work, however, is more about the query-click graph and its properties. There has also been a lot of work on query clustering by common intent using this graph, e.g., Yi and Maghoul (2009) and Wen et al. (2002). We focus not on clustering but on understanding the expression of intent.

3 Method

We address the main thesis of the work by retrospectively studying queries issued to a search engine over the course of 12 consecutive months.

$\mathcal{Q}$-queries. First we define a notion of question queries based on the standard definition of questions in English. A query is a *$\mathcal{Q}$-query* if it contains at least two tokens and satisfies one of the following criteria.

(i) Starts with one of the interrogative words, or $\mathcal{Q}$-words ("*how, what, which, why, where, when, who, whose*").

(ii) Starts with "*do, does, did, can, could, has, have, is, was, are, were, should*". While this ensures a legitimate question in well-formed English texts, in queries, we may get "*do not call list*". Thus, we insist that the second token cannot be "*not*".

(iii) Ends with a question mark ("*?*").

Otherwise it is a *$\overline{\mathcal{Q}}$-query*. The list of keywords ($\mathcal{Q}$-words) is chosen using an English lexicon. Words such as "*shall*" and "*will*", even though interrogative in nature, introduce more ambiguity (e.g., "*shall we dance lyrics*" or "*will smith*") and do not account for much traffic in general; discarding such words will not impact the findings.

Co-click data on "stable" URLs. We work with the set of queries collected between Dec 2009 and Nov 2010 from the Yahoo! querylog. We gradually refine this raw data to study changes in query formulation over comparable and consistent search intents.

1. $\boldsymbol{S}_{\textbf{all}}$ consists of all incoming search queries after preprocessing: browser cookies[2] that correspond to possible robots/automated queries and queries with non-alphanumeric characters are discarded; all punctuations, with the exception of "*?*", are removed; all remaining tokens are lower-cased, with the original word ordering preserved.

2. $\boldsymbol{C}_{\textbf{all}}$ consists of queries formulated for similar *search intent*, where intent was approximated by the result URL clicked in response to the query. That is, we assume queries that lead to a click on the same URL are issued with similar information need. To reduce the noise introduced by this approximation when users explore beyond their original intent, we focus on (query, URL) pairs where the URL u was clicked from top-10 search results[3] for query q.

3. $\boldsymbol{U}_{\boldsymbol{\mathcal{Q}}}^{\mathbf{c50}}$ is our final dataset with queries grouped over "stable" intents. First, for each month m, we collect the multiset C_i of all (q, u_i) pairs for each clicked URL u_i, where the size of C_i is the total number of clicks received by u_i during m. Let

[2] We approximate user identity via the *browser cookie* (which are anonymized for privacy). While browser cookies can be unreliable (e.g, they can be cleared), in practice, they are the best proxy for unique users.

[3] In any case, clicks beyond top-10 results (i.e., the first result page) only account for a small fraction of click traffic.

$U^{(m)}$ be all URLs for month m. We restrict to $U = \bigcap_m U^{(m)}$. This set represents intents and contents that persist over the 12-month period, allowing us to examine query formulation changes over time.

We then extract a subset U_Q of U consisting of the URLs associated with at least one $\mathcal{Q}$-query in one of the months. Interestingly, we observe that $\frac{|U_Q|}{|U|} = 0.55$: roughly half of the "stable" URLs are associated with at least one $\mathcal{Q}$-query!

Finally, we restrict to URLs with at least 50 clicks in each month to obtain reliable statistics later on. U_Q^{c50} consists of a random sample of such URLs, with 423,672 unique URLs and 231M unique queries (of which 21M (9%) are $\mathcal{Q}$-queries).

$\mathcal{Q}$-level. For each search intent (i.e., a click on u), to capture the degree to which people express that intent via $\mathcal{Q}$-queries, we define its *$\mathcal{Q}$-level* as the fraction of clicks on u from $\mathcal{Q}$-queries. Since we are interested in general query formulation behavior, we do not want our analysis to be dominated by trends in popular intents. Thus, we take macro-average of $\mathcal{Q}$-level over different URLs in a given month, and our main aim is to explore long-term temporal changes in this value.

4 Results

4.1 Characteristics of $\mathcal{Q}$-queries

Are $\mathcal{Q}$-queries really questions? We examine 100 random queries from the least frequent $\mathcal{Q}$-queries in our dataset. Only two are false-positives: "*who wants to be a millionaire game*" (TV show-based game) and "*can tho nail florida*" (a local business). The rest are indeed question-like: while they are not necessarily grammatical, the desire to express the intent by posing it as a question is unmistakable.

Still, are they mostly ostensible questions like "*how find network key*", or well-formed full-length questions like "*where can i watch one tree hill season 7 episode 2*"? (Both are present in our dataset.)

Given the lack of syntactic parsers that are appropriate for search queries, we address this question using a more robust measure: the probability mass of *function words*. In contrast to content words (open class words), function words (closed class words) have little lexical meaning — they mainly provide grammatical information and are defined by their syntactic behavior. As a result, most function words are treated as stopwords in IR systems, and web users often exclude them from queries. A high fraction of function words is a signal of queries behaving more like normal texts in terms of the amount of tokens "spent" to be structurally complete.

We use the list of function words from Sequence Publishing[4], and augment the auxiliary verbs with a list from Wikipedia[5]. Since most of the $\mathcal{Q}$-words used to identify $\mathcal{Q}$-queries are function words themselves, a higher fraction of function words in $\mathcal{Q}$-queries is immediate. We remove the word used for $\mathcal{Q}$-query identification from the input string to avoid trivial observations. That is, "*how find network key*" becomes "*find network key*", with zero contribution to the probability mass of function words.

The following table summarizes the probability mass of function words in all unique $\overline{\mathcal{Q}}$-queries and $\mathcal{Q}$-queries in U_Q^{c50}, compared to two natural-language corpora: a sample of 6.6M questions posted by web users on a community-based question-answering site, Yahoo! Answers ($Q_{\text{Y!A}}$), and the Brown corpus[6] (Br). All datasets went through the same query preprocessing steps, as well as the $\mathcal{Q}$-word-removal step described above.

Type	$\overline{\mathcal{Q}}$-q	$\mathcal{Q}$-q	$Q_{\text{Y!A}}$	Br
Auxiliary verbs	0.4	8.5	8.1	5.8
Conjunctions	1.2	1.4	3.4	4.5
Determiners	2.0	8.7	8.2	10.1
Prepositions	6.5	13.7	10.1	13.3
Pronouns	0.7	3.4	9.1	5.9
Quantifiers	0.1	0.7	0.4	0.6
Ambiguous	2.1	2.7	4.6	7.0
Total	12.9	39.0	43.9	47.1

Clearly, $\mathcal{Q}$-queries are more similar to the two natural-language corpora in terms of this shallow measure of structural completeness. Notably, they contain a much higher fraction of function words compared to $\overline{\mathcal{Q}}$-queries, even though they express similar search intent.

This trend is consistent when we break down by type, except that $\mathcal{Q}$-queries contain fewer conjunctions and pronouns compared to $Q_{\text{Y!A}}$ and Br. This happens since $\mathcal{Q}$-queries do not tend to have complex sentence or discourse structures. Our results

[4] www.sequencepublishing.com/academic.html.

[5] en.wikipedia.org/wiki/List_of_English_auxiliary_verbs

[6] khnt.aksis.uib.no/icame/manuals/brown/

suggest that if users express their information need in a question form, they are more likely to express it in a structurally complete fashion.

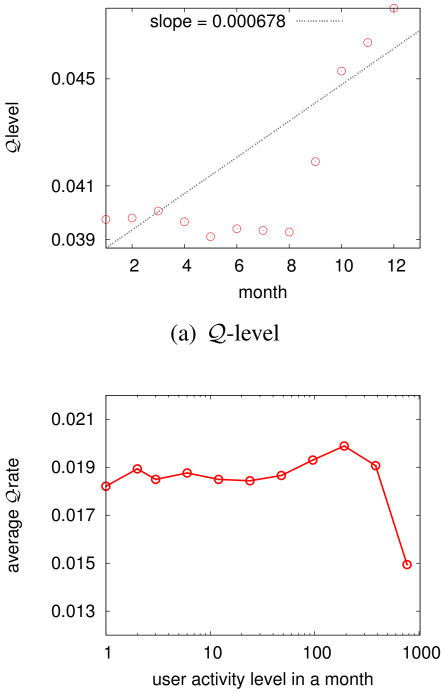
Lastly, we examine the length of Q-queries and $\overline{Q}$-queries in each multiset C_i. If $\overline{Q}$-queries contain other content words in place of Q-words to express similar intent (e.g., "*steps to publish a book*" vs. "*how to publish a book*"), we should observe a similar length distribution. Instead, we find that on average Q-queries tend to be longer than $\overline{Q}$-queries by 3.58 tokens. Even if we remove the Q-word and a companion function word, Q-queries would still be one to two words longer. In web search, where the overall query traffic averages at shorter than 3 tokens, this is a significant difference in length — apparently people are more generous with words when they write in the question mode.

4.2 Trend of Q-level

We have just confirmed that Q-queries resemble natural-language questions to a certain degree. Next we turn to our central question: how does Q-level (macro-averaged over different intents) change over time? To this end, we compute a linear regression of Q-level across 12 months, conduct a hypothesis test (with the null hypothesis being the slope of the regression equal to zero), and report the P-value for two-tailed t-test.

As shown in Figure 1(a), there is a mid-range correlation between Q-level and time in U_Q^{c50} (correlation coefficient $r = 0.78$). While the trend is measured with slope $=$ 0.000678 (it would be surprising if the slope for the average behavior of this many users were any steeper!), it is statistically significant that Q-level is growing over time: the null hypothesis is rejected with $P < 0.001$. That is, over a large collection of intents and contents, users are becoming more likely to formulate queries in question forms, even though such content could easily be reached via non-question-queries.

One may question if this is an artifact of using "stable" clicked URLs. Could it be that search engines learn from user behavior data and gradually present such URLs in lower ranks (i.e., shown earlier in the page; e.g., first result returned), which increases the chance of them being seen and clicked? This is indeed true, but it holds for both Q-queries and $\overline{Q}$-queries. More specifically, if we consider the rank of the clicked URL as a measure of search result quality (the lower the better), we observe improvements for both Q-queries and $\overline{Q}$-queries over time (and the gap is shortening). However, the average click position for Q-queries is consistently higher in rank throughout the time. Thus, it is not because the search engine is answering the Q-queries better than $\overline{Q}$-queries that users start to use Q-queries more. While we might still postulate that the decreasing gap in search quality (as measured by click positions) might have contributed to the increase in Q-level, if we examine the co-click data without the stability constraint, we observe the following: an increasing click traffic from Q-queries and an increasing gap in click positions between Q-queries and $\overline{Q}$-queries.



(a) Q-level

(b) Q-rate

Figure 1: Q-level for different months in U_Q^{c50}; Q-rate for users with different activity levels in S_{all}.

In addition, we also observe an upward trend for the overall incoming query traffic accounted for by Q-queries in S_{all} (slope = 0.000142, $r = 0.618$, $P < 0.05$). The upward trend in the fraction of unique queries coming from Q-queries is even more pronounced (slope = 0.000626, $r = 0.888$, $P < 0.001$). While this trend could be partly due to differences in search intent, it nonetheless reinforces the general message of increases in $\mathcal{Q}$-queries usage. This is also consistent with the anecdotal evidence from Google trends (Section 1) suggesting that the trends we observe are not search-engine specific and have been in existence for over a year.[7]

4.3 Observations in the overall query traffic

Note that in $U_{\mathcal{Q}}^{c50}$, $\mathcal{Q}$-level averages $\sim 4\%$; recall also for a rather significant portion of the web content, at least one user chose to formulate his/her intent in $\mathcal{Q}$-queries ($\frac{|U_{\mathcal{Q}}|}{|U|} = 0.55$). Both reflect the prevalence of $\mathcal{Q}$-queries. Is that specific to well-constrained datasets like $U_{\mathcal{Q}}^{c50}$? We examine the overall incoming queries represented in S_{all}. On average, $\mathcal{Q}$-queries account for 1.8% of query traffic. 5.7% of all unique queries are $\mathcal{Q}$-queries, indicating greater diversity in $\mathcal{Q}$-queries.

What types of questions do users ask? The table below shows the top $\mathcal{Q}$-words in the query traffic; "*how*" and "*what*" lead the chart.

word	%	word	%	word	%
how	0.7444	*what*	0.4360	*where*	0.0928
?	0.0715	*who*	0.0684	*is*	0.0676
can	0.0658	*why*	0.0648	*when*	0.0549
do	0.0295	*does*	0.0294	*are*	0.0193
which	0.0172	*did*	0.0075	*should*	0.0072

How does the query traffic associated with different $\mathcal{Q}$-words change over time? We observe that all slopes are positive (though not all are statistically significant), indicating that the increase in $\mathcal{Q}$-queries happens for different types of questions.

Is it only a small number of amateur users who persist with $\mathcal{Q}$-queries? We define *$\mathcal{Q}$-rate* for a given user (approximated by browser cookie b) as the fraction of query traffic accounted for by $\mathcal{Q}$-queries. We plot this against b's activity level, measured by the number of queries issued by b in a month. We binned users by their activity levels on the $\log_2$-scale and compute the average $\mathcal{Q}$-rate for that bin. As shown in Figure 1(b), relatively light users who issue up to 30 queries per month do not differ much in $\mathcal{Q}$-rate on an aggregate level. Interestingly, mid-range users (around 300 queries per month) exhibit higher $\mathcal{Q}$-rate than the light users. And for the most heavy users, the $\mathcal{Q}$-rate tapers down.

Furthermore, taking the data from the last month in S_{all}, we observe that for users who issued at least 258 queries, more than half of them have issued at least one $\mathcal{Q}$-query in that month — using $\mathcal{Q}$-queries is rather prevalent among non-amateur users.

5 Concluding remarks

In this paper we study the prevalence and characteristics of natural-language questions in web search queries. To the best of our knowledge, this is the first study of such kind. Our study shows that questions in web search queries are both prevalent and temporally increasing. Our central observation is that this trend holds in terms of how people formulate queries for the same search intent (in the carefully constructed dataset $U_{\mathcal{Q}}^{c50}$). The message is reinforced as we observe a similar trend in the percentage of overall incoming query traffic being $\mathcal{Q}$-queries; in addition, anectodal evidence can be obtained from Google trends.

We recall the following two findings from our study. (a) Given the construction of $U_{\mathcal{Q}}^{c50}$, the upward trend we observe is not a direct result of users looking for different types of information, although it is possible that the rise of Q&A sites and users entrusting search engines with more complex information needs could have indirect influences. (b) The results in Section 4.2 suggest that in $U_{\mathcal{Q}}^{c50}$, $\mathcal{Q}$-queries receive inferior results than $\overline{\mathcal{Q}}$-queries (i.e., higher average rank for clicked results for $\mathcal{Q}$-queries for similar search intents), thus the rise in the use of $\mathcal{Q}$-queries is not a direct result of users learning the most effective query formulation for the search engine. These suggest an interesting research question: what is causing the rise in question-query usage?

Irrespective of the cause, given that there is an increased use of $\mathcal{Q}$-queries in spite of the seemingly inferior search results, there is a strong need for the search engines to improve their handling of question-queries.

Acknowledgments

We thank Evgeniy Gabrilovich, Lillian Lee, D. Sivakumar, and the anonymous reviewers for many useful suggestions.

[7] An explanation of why the upward trend starts at the end of 2007 is beyond the scope of this work; we postulate that this coincides with the rise in popularity of community-based Q&A sites.

References

Anne Aula, Rehan M. Khan, and Zhiwei Guan. 2010. How does search behavior change as search becomes more difficult? In *Proc. 28th CHI*, pages 35–44.

Ricardo Baeza-Yates and Alessandro Tiberi. 2007. Extracting semantic relations from query logs. In *Proc. 13th KDD*, pages 76–85.

Cory Barr, Rosie Jones, and Moira Regelson. 2008. The linguistic structure of English web-search queries. In *Proc. EMNLP*, pages 1021–1030.

Steven M. Beitzel, Eric C. Jensen, Abdur Chowdhury, David Grossman, and Ophir Frieder. 2004. Hourly analysis of a very large topically categorized web query log. In *Proc. 27th SIGIR*, pages 321–328.

M. Bendersky and W. B. Croft. 2009. Analysis of long queries in a large scale search log. In *Proc. WSDM Workshop on Web Search Click Data*.

Susan Dumais, Michele Banko, Eric Brill, Jimmy Lin, and Andrew Ng. 2002. Web question answering: Is more always better? In *Proc. 25th SIGIR*, pages 291–298.

Mark Kröll and Markus Strohmaier. 2009. Analyzing human intentions in natural language text. In *Proc. 5th K-CAP*, pages 197–198.

Cody Kwok, Oren Etzioni, and Daniel S. Weld. 2001. Scaling question answering to the web. *ACM TOIS*, 19:242–262.

Josiane Mothe and Ludovic Tanguy. 2005. Linguistic features to predict query difficulty. In *Proc. SIGIR Workshop on Predicting Query Difficulty - Methods and Applications*.

Marius Pasca. 2007. Weakly-supervised discovery of named entities using web search queries. In *Proc. 16th CIKM*, pages 683–690.

Fabrizio Silvestri. 2010. Mining Query Logs: Turning Search Usage Data into Knowledge. *Foundations and Trends in Information Retrieval*, 4(1):1–174.

Amanda Spink and H. Cenk Ozmultu. 2002. Characteristics of question format web queries: An exploratory study. *Information Processing and Management*, 38(4):453–471.

Markus Strohmaier and Mark Kröll. 2009. Studying databases of intentions: do search query logs capture knowledge about common human goals? In *Proc. 5th K-CAP*, pages 89–96.

Ji-Rong Wen, Jian-Yun Nie, and Hong-Jiang Zhang. 2002. Query clustering using user logs. *ACM TOIS*, 20:59–81.

Jeonghee Yi and Farzin Maghoul. 2009. Query clustering using click-through graph. In *Proc. 18th WWW*, pages 1055–1056.