What Users Ask a Search Engine: Analyzing One Billion Russian Question Queries

Michael Völske
Bauhaus-Universität Weimar
michael.voelske@uni-weimar.de

Pavel Braslavski
Ural Federal University
pbras@yandex.ru

Matthias Hagen
Bauhaus-Universität Weimar
matthias.hagen@uni-weimar.de

Galina Lezina
Ural Federal University
galina.lezina@gmail.com

Benno Stein
Bauhaus-Universität Weimar
benno.stein@uni-weimar.de

ABSTRACT

We analyze the question queries submitted to a large commercial web search engine to get insights about what people ask, and to better tailor the search results to the users' needs. Based on a dataset of about one billion question queries submitted during the year 2012, we investigate askers' querying behavior with the support of automatic query categorization. While the importance of question queries is likely to increase, at present they only make up 3–4% of the total search traffic.

Since questions are such a small part of the query stream, and are more likely to be unique than shorter queries, click-through information is typically rather sparse. Thus, query categorization methods based on the categories of clicked web documents do not work well for questions. As an alternative, we propose a robust question query classification method that uses the labeled questions from a large community question answering platform (CQA) as a training set. The resulting classifier is then transferred to the web search questions. Even though questions on CQA platforms tend to be different to web search questions, our categorization method proves competitive with strong baselines with respect to classification accuracy.

To show the scalability of our proposed method we apply the classifiers to about one billion question queries and discuss the trade-offs between performance and accuracy that different classification models offer. Our findings reveal what people ask a search engine and also how this contrasts behavior on a CQA platform.

Categories and Subject Descriptors: H.3.3 [Information Systems]: Information Search and Retrieval

Keywords: Question Queries; Query Log Analysis; Query Classification; Community Question Answering (CQA).

CIKM'15, October 19–23, 2015, Melbourne, Australia.

ACM 978-1-4503-3794-6/15/10 ...$15.00.
DOI: http://dx.doi.org/10.1145/2806416.2806457

1. INTRODUCTION

Questions are a natural form of expressing information needs. People ask questions when they seek information, help, or advice. Web search engines have taught users the "telegram style" of keyword search queries such as `[lose weight]`. Nevertheless, the share of natural language questions, for example `[how much exercise should i do to lose 10 pounds]`, in search query logs is increasing [21].

In the late 1990s, queries in question form comprised less than 1% of the query stream of a general-purpose search engine; the most common format was `[where can i find ...]` for general information on a topic [28]. Pang and Kumar report that question queries accounted for about 2% of the entire Yahoo query stream in 2010 [21]. Our analysis shows that question queries already constitute a 3–4% -share of the query log we are using from 2012; questions are thus still on the rise even in keyword-based interfaces.

But why do users formulate search queries as questions? A possible explanation is the general tendency of smoother, more natural human-computer interaction with information retrieval systems, as evidenced by touch, voice, and visual search interfaces. Although voice search queries constitute a marginal share of the entire query log, they support this trend [25]. The spread of community question answering (CQA) services such as Yahoo! Answers provides a parallel setting in which to study question-asking behavior on the web.[1] CQA sites allow users to pose questions to other community members, to answer questions, rate questions and answers, and receive feedback. The services are quite popular and have collected a vast amount of content in the form of questions and answers that is being indexed by major search engines. The odds are high that a question a user has in mind has already been asked by someone before and can be found through search engines.

However, submitting queries in the form of natural language questions does not always yield better search results. As several studies show [7, 1, 21], web search engines perform worse at answering question queries compared to corresponding keyword queries. In view of the growing share of question queries, and the still lagging search quality for them, there is a strong need to improve the processing of such queries. For example, Google's latest major search algorithm update

[1] `https://answers.yahoo.com/`

in the fall of 2013—codenamed Hummingbird—was targeted at answering long natural questions better.[2]

One option to improve the search results for question queries is accounting not only for the query's terms, but also its topic, by means of query classification. The technique has benefited general search [2], query disambiguation and routing to vertical search [16], and search advertising [9].

The main difficulty in query classification is data sparseness: the short query strings. Search queries contain around three words on average [22, 7], and despite the fact that question queries are somewhat longer—around six to seven words, according to different studies [21, 19]—they are still much shorter than web documents.

Our approach exploits CQA data and its categorization scheme as a "bridge classification" for the question query classification problem. CQA services provide a vast amount of questions manually categorized by their users that can inform automatic query categorization. Similarly to [16], our primary goal is to expand the training set, using rather straightforward classification techniques. In our study we employ several million user-generated questions, along with top-level category labels, for building a question-query classifier. To the best of our knowledge, this approach is novel.

Robust topical classification can also boost the identification of users' information needs in contexts different from web search. Mobile voice-activated assistants like Apple's Siri—that suffer from a very limited range of available classification domains [6]—may benefit just as the analysis of short interrogative posts on Twitter [33] or Facebook [20].

Our contributions are two-fold. First, we describe and analyze two large complementary datasets of Russian questions from 2012: (1) a year's worth of questions posted at a popular CQA service, and (2) question queries submitted to a large commercial search engine. To the best of our knowledge, this is the first study dealing with non-English question datasets of this size. Second, we build a question classifier of high quality using CQA data and use it to analyze the information needs of web search question askers.

The paper is organized as follows: In Section 2, we discuss the literature on query classification and question analysis. We then introduce the datasets used in our analyses in Section 3 and explain our classification approach in Section 4. Besides experimental evaluation of the classification approaches, Section 5 also shows the application of our classification approach to one billion web search question queries to shed some light on what people ask their search engine. Finally, Section 6 summarizes the results and suggests interesting directions for future work.

2. RELATED WORK

Question queries have been the subject of dedicated search log studies [28] and have been analyzed in the context of long queries [7]. Pang and Kumar [21] draw attention to the phenomenon of question queries in search engine logs, describe their structural and statistical characteristics, and show that the share of these queries grows. A more recent longitudinal study on the evolution of user behavior shows questions as an important part [17]. The authors also note that the search results for question queries are usually worse than for the corresponding keyword queries expressing the same information need [7, 1, 21].

[2]http://onforb.es/1bfagwI

The data sparseness problem is usually addressed by enriching queries with additional information. Queries are categorized based on the category labels of documents returned by a search engine [9] or are enriched by the search results containing document titles and snippets [26]. Bailey and coauthors [2] classify long queries with sparse user interaction data by matching them against shorter and more popular queries categorized based on past users' behavior. Li et al. [16] suggest to substantially expand the set of labeled queries using click-through information: user clicks on the same link returned for different queries are considered as a similarity indicator. Thus, iterative propagation of category labels from seed queries along click edges through co-clicked documents to unlabeled queries allows expanding the initial training set by several orders of magnitude.

Note that, to be practically applicable, all three approaches require the availability of search log information. In case of click-through information this is rather obvious. In case of using returned results or titles and snippets for categorization, the classification can be accounted for in a second retrieval run or performed off-line and then applied on-the-fly if the query appears again. In case of questions, however, the availability of click-through data is a big problem, as questions are typically rather unique, and have little associated log data. This rules out the above classification methods for our use case of question classification and we aim for another approach to analyze our large question query log.

In contrast to classifying isolated queries, topical categorization of a large log should give high-level insights into the spectrum of user interests and their dynamics. Spink et al. [29] manually label several thousand queries from a search log in an attempt to study user interests. Later, Beitzel et al. [4] automatically match queries against manually compiled topical word lists, classifying 13% of a search engine's query stream. A bootstrapping of this method based on word-category distributions yields an improved recall [5] but still low coverage. In a fully automatic large-scale analysis, Bar-Ilan et al. [3] perform a topical classification of the AOL and MSR logs using an SVM classifier over query word uni- and bigrams. Similarly, one of our methods uses query word unigrams for classification.

Besides the large-scale analysis, another important aspect of our study is that we conduct experiments on log data spanning one year. Up to today there are only few studies dealing with query data stretching over such long periods. Richardson [24] explores the long-term dependencies of users' intents and preferences based on a one-year log of millions of users. He concludes that the analysis of user behavior based on such long periods can uncover information not present in shorter logs, and as such be of interest not only for information retrieval, but also for social sciences, psychology, market research, and medical studies. Note that in contrast to Richardson's study, we aim at analyzing question queries in particular, but still the dimensions of the employed log data are comparable. The aforementioned work by Pang and Kumar [21] also draws conclusions based on the analysis of an annual search log, but other studies used much smaller logs. Beitzel et al. [4] explore the topical structure of a six month query log and Liu et al. [17] track user behavior based on log excerpts spanning two weeks in three subsequent years.

There are several studies conducted on the intersection of web search and CQA. Weber et al. [30] aim at finding answers (tips) to web queries with *how-to* intent (not necessarily ex-

pressed as well-formed *how-to* questions) in Yahoo! Answers archives. Liu et al. [18] evaluate the utility of existing CQA answers in web search scenarios. In a follow-up study [19] the authors track users, who follow up web searching with asking a question on a CQA platform. In contrast to these studies, our goal is not to develop methods that provide better answers to questions or that recommend CQA items to the users. Instead, we aim at the topical categorization of questions on the scale of a year; the results might then improve retrieval systems, as is proposed in some studies [14, 11].

Topical categorization of questions posted on CQA services is the subject of several studies. For instance, Li et al. [15] suggest to use topic information in a question routing task (i.e., delivering newly posted questions to potential answerers). While this use case is rather different from ours, the study of Qu et al. [23] who investigate the contribution of different components to question classification quality (machine learning methods, n-gram features, data fields, and training sample size) is closer to our setting. We incorporate several of their findings in one of our methods using bag-of-words features. Chan et al. [12] apply a set of kernels corresponding to different aspects of questions to hierarchical question classification. Since we are interested in a rather broad, non-nested category scheme that can be used in the actual retrieval process, we do not aim for any hierarchy. Cai et al. [10] propose to enrich CQA questions with Wikipedia entries as a means to counter the sparseness problem discussed above. However, in contrast to our approach, all these CQA methods do not extend beyond CQA (i.e., they use CQA data for learning and consequently perform classification on the data of the same origin). One of our contributions instead is to show how a classifier trained on CQA questions can be used to classify web search questions as well, affording the opportunity of including on-the-fly class information in the retrieval process for questions submitted to search engines.

Question analysis in other domains, such as questions posted on Twitter [33] or Facebook [20], may also benefit from such an online classification method.

3. DATASETS

The basis for our question query classification are two datasets: a large amount of question-like queries collected from the query log of Yandex[3], a leading Russian search engine, and a year's worth of questions and answers from a popular Russian community question answering (CQA) platform Otvety@Mail.Ru[4]. Both datasets contain Russian queries only, although some of the queries contain words in other languages (mainly named entities such as movie or song titles, video games names, etc.). Below, we outline the data acquisition process and provide further details on the datasets.

3.1 Web Search Questions

The initial dataset comprises of all queries from Yandex logs for the year 2012 containing one of 58 combinations of question word uni- or bigrams (e.g., *what, where, when, why, how, does, should, . . . , in which, for what*, etc.). This is similar to previous processes of question extraction from query logs [7] except that the question word set was adapted to Russian. Each entry in the resulting question excerpt is annotated with the query string, a time stamp, and user ID. The nearly 2 billion initially acquired questions form about 3–4% of the actual query log, indicating some further increase in the number of questions submitted to web search engines compared to the 2010 Yahoo figure of about 2% using similar extraction rules [21]. Under the agreement with the search engine, we have access only to the queries containing question words for research purposes; we have no access to the other queries issued by the same users or to the search results. Since it was curated at the end of 2012, the query log contains no entries for the second half of December. Hence, we omit all December entries from our analysis.

Cleaning step	Unique users	Questions
Raw log	185,700,840	1,980,878,942
Spam & bots	184,630,648	1,903,716,272
Core questions	167,812,003	1,577,657,443
Repeats & prefixes	167,812,003	1,265,433,864
Unoriginal questions	145,688,746	923,482,955
Single-word questions	145,071,912	915,055,325

Table 1: Cleaning the question queries extracted from the web query log.

In an iterative process outlined below, we apply further cleaning steps to keep only queries that represent actual question-asking information needs. Table 1 shows the individual steps of the data cleaning process and their impact.

In a first step, we remove spam and bot queries from the log. After examining user activity statistics, we suggest to characterize a user as a bot when any of the following properties holds: (1) more than 2,000 total interactions over the entire year; (2) more than five questions within the most active one-minute window; (3) a median question length of more than 20 words; or (4) at least 50 questions in total, and the same leading 15 characters in at least 80% of them.

The first two criteria are aimed at the number of questions per time slot, while the latter two are aimed at the type of questions submitted. Users submitting a very large number of questions in one year or in their peak activity minute behave rather "unhuman" and we view them as bots. Users submitting unusually long questions, or questions almost always starting with the same 2–3 words are also behaving rather unnaturally. Extensive spot-check inspections of users matching any of the above four criteria showed that all of them could easily be viewed as bots. The specific numbers might be debatable, especially for the peak activity for some of the affected users, but we decided to rather aggressively remove users to base later examinations only on questions that were very likely submitted as a human information need.

Altogether, the first cleaning step removed about 1 million users and all their 77 million questions. Examples of removed bots include users submitting very many `[how to translate ...]` or `[how is the weather in ...]` questions that probably aim at scraping the search engine's translation or weather service, or for instance bots submitting thousands of long copy-pasted questions from exams. Interestingly, hardly any of the questions containing an actual question mark remain after the first cleaning step; the ones that do remain almost always also seem to be copy-pasted from some exam. Having removed all entries for these suspicious users, we apply subsequent filtering steps to individual questions in the log.

In a second cleaning step, we retain only "core questions" with a question word in the first position, since extensive

[3] `http://yandex.ru`
[4] `http://otvet.mail.ru`

spot checks of the other queries showed a large number of queries with debatable question intent. Instead of devising sophisticated rules to decide for each such query whether it actually is a question intent or not, we again choose an aggressive removal to reduce the amount of non-question needs in the final dataset. This step removes about 326 million questions, and about 17 million users that would not have any remaining question.

The third cleaning step eliminates repeated questions and collapses prefixes. The goal is to remove bogus query submissions resulting from instant search, accidental submissions of unfinished question strings, or log entries of users paging through search engine results pages (SERPs) (always with the same question string but not really submitting new queries). If a user resubmits the same question within 90 minutes, without a different question in between, we only retain the first occurrence. To catch SERP paging behavior, we again choose to aggressively clean the query log using a long temporal window rather than a 2- or 5-minute gap. To remove "unfinished" questions stemming from instant searches or unintentionally submitted queries, we analyze pairs of questions submitted within 5 seconds. When the first query of such a pair is a prefix of the second, we retain only the latter (e.g., `[when was caesar bo]` is removed when immediately followed by `[when was caesar born]`).

In a fourth cleaning step, we remove unoriginal questions, by which we refer to questions not formulated by the user themselves, but probably stemming from some external source. For this, we first remove all questions that match one of 885 titles of Wikipedia articles (e.g., the movie title `[what women want]`). We then also identify some questions that seek answers to crossword puzzles; we assume a crossword puzzle information need if the query ends with the phrase `[`n `words]` for some value of n. In addition, we also remove question queries that contain the phrase `[family feud]`, and variants of that TV show's name in its Russian incarnation.[5] For both the crossword and TV show questions, we include a "bootstrapping" step, in which we also remove all the questions that co-occur ten or more times with one of the characteristic phrases (about 7,600 question strings identified in the bootstrapping). Furthermore, we also remove questions matching a list of 1,764 questions published on fan websites of the Family Feud show. We believe that hardly any of the questions matching a Wikipedia article with the very same title, a crossword puzzle question, or a Family Feud question actually represent an original question intent of the user.

Finally, in a fifth cleaning step, we filter out those question queries that contain only one word after stopword and question word removal. Although this also removes questions like `[when is christmas]` our spot checks showed many of the single-word questions not to represent real question needs.

The cleaning steps removed more than half of the originally sampled questions; the remaining dataset contains about 915 million question queries from about 145 million users. This represents about 1–2% of the search engine's query stream (cf. Figure 1 for the monthly fraction). Further characteristics and a comparison to our CQA dataset can be found in Section 3.4.

[5]Family Feud is a popular TV show that prominently features questions like `[what is a problem most people have in their life]` for which the participants have to guess the most popular response of 100 people being asked that question.

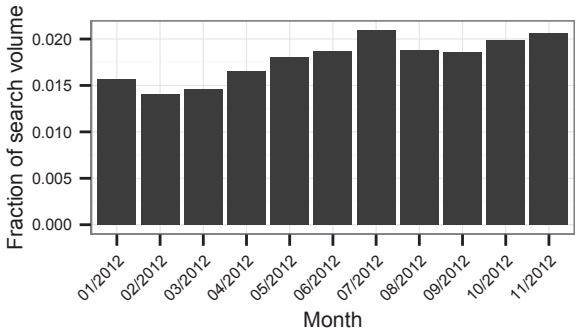

Figure 1: Question queries in the cleaned dataset as a monthly fraction of the total query traffic.

3.2 Community Question Answering Data

The CQA dataset we acquired comprises approximately 11 million questions submitted in Russian by over 2 million unique users to the Russian CQA platform Otvety@Mail.Ru[6] throughout the year 2012. Otvety@Mail.Ru (*otvety* means *answers*) is a Russian counterpart of Yahoo! Answers with similar rules and incentives. Each question is manually categorized by the submitter into one of 28 top-level categories with altogether 189 leaf-level categories forming a two-level hierarchy. In the process of dataset acquisition, we omit several ambiguous categories, and merge closely related categories, leaving the 14 top-level categories shown in the first column of Table 2 as our classification targets.

When using query category labels as additional features for ranking along hundreds of other features, coarse-grained flat categories usually suffice. This is an important difference to query classification for search advertising (advertising-to-query matching is based on category information only), automatic classification of web documents, or category suggestion for questions in the CQA scenario. In both latter cases, the amount of information items under leaf categories must be "digestible" by humans. Hence, hierarchical taxonomies with thousands of categories are used.

We paid special attention to noise in category labels, dissimilarity of the topic distributions in the two datasets, and the alignment of source (CQA) and target categories. Since the user posting a question on the CQA platform manually labels the question with a category—and this seems to be an error-prone task given the number of categories—we decided to further clean the initial dataset. We only keep questions submitted by users that have posted at least three questions that got an answer. This criterion is meant to capture questions with better categorizations. Users posting more than just one or two test questions can be viewed as more experienced with the category scheme and questions that got an answer form a further support of this hypothesis since other users found the query under its category.

The assigned categories in the remaining 6 million questions from the CQA platform are less noisy than the original 11 million questions, making the cleaned CQA data well-suited as a training set for our query classification task. The second column of Table 2 shows the number of instances in

[6]`http://otvet.mail.ru/`

Category	Number of instances	
	CQA	Test set
Society & Culture	1,267,700	95
Computers & Internet	965,834	131
Family & Relationships	950,180	33
Adult	526,465	13
Games & Recreation	524,533	61
Education	372,600	38
Home & Garden	355,906	117
Entertainment & Music	337,364	64
Cars & Transportation	335,659	89
Health	307,033	70
Consumer Electronics	193,685	43
Beauty & Style	173,825	23
Sports	165,959	16
Business & Finance	99,524	41
Σ	6,576,267	834

Table 2: Class distribution in the CQA dataset and the manually labeled question query test set.

the CQA dataset per category. Further characteristics in comparison to our question queries dataset can be found in Section 3.4.

3.3 Web Search Question Test Data

In order to evaluate the performance of our classification pipeline on the question queries from the search engine log, we randomly sample 1,000 entries from the cleaned dataset. After labeling by three domain experts, no two annotators picked the same category for 166 of the questions. These more ambiguous questions were removed from the test set. The third column of Table 2 shows the class distribution in the remaining test set of 834 questions.

3.4 Descriptive Statistics

We have about 915 million questions from about 145 million users in the web search question query dataset and about 6 million questions from about 0.5 million users in the CQA dataset. The distribution of the number of questions per user per dataset is shown in Figure 2.

Note that we do not have users with less than three questions in the CQA dataset due to our filtering rule. About one third of the CQA users in our cleaned dataset have posted three questions, another half have posted at most

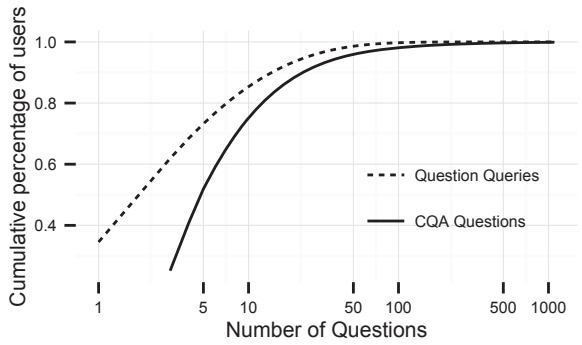

Figure 2: Number of questions per user.

N-gram	English translation	Frequency
CQA Dataset		
можно ли	whether is it possible	113,291
а вы	and you	89,193
у меня	I	68,337
что делать	what to do	66,665
как вы	how you	64,235
что делать если	what to do if	36,990
где можно скачать	where can I download	21,665
как вы думаете	what do you think	20,057
а у вас	and you	16,631
как вы относитесь	what do you think	13,497
Question Queries Dataset		
как сделать	how to make	35,678,293
можно ли	whether is it possible	28,001,988
как правильно	how correctly	23,014,202
сколько стоит	how much costs	19,533,978
где купить	where to buy	11,405,702
как избавиться от	how to get rid of	5,166,515
где можно купить	where to buy	2,804,874
как скачать музыку	how to download music	2,072,003
как доехать до	how to get to	2,028,746
какие документы нужны	what documents are needed	1,818,986

Table 3: The five most frequent initial 2- and 3-grams per dataset.

ten questions and the remaining 20% have submitted up to 5,000 questions in the year 2012. The average number of questions per user in the CQA data is about 16, with a maximum of 257 questions.

In the question queries data, the situation is slightly different, with the average user submitting about 6 questions; this is not surprising since we did not remove users with very few questions here. About 40% of the users only submitted a single question in the whole year. However, due to the user identification method on the server side, some questions from the same user might get logged with different user IDs. Similarly to the CQA data, another 40–50% of the users submit at most ten questions while only 10% of the users submit up to 2,000 questions in the whole year. Since 2,000 questions per year was a bot-removal threshold used in our cleaning process, there are no users with more than 2,000 question queries and only a few with more than 1,000 questions; the most interrogative user submitted about 1,500 question queries in the whole year.

The two datasets also differ in the most frequent question prefixes given in Table 3. Not surprisingly, the top prefixes of the question titles in the CQA data show that users often do not explicitly formulate a question but rather ask others for help (e.g., `[I need your help]` or `[can you help me]`). Due to the sampling strategy, the question queries have explicit question words as their initial n-grams. As was already observed in other studies, the most frequent questions are how-to questions that can be formulated using different bi- and trigrams in Russian.

Figure 3 shows the datasets' frequency distributions. In the CQA data, about 98% of queries are unique, as opposed to 88% for the question queries. Not surprising, given that we count title and description as the query, the average question appears just once in the CQA data with the very short most frequent questions appearing five times. In the search engine question queries, the average question appears about two times while the most frequent question, `[how to download music from VK]`, has nearly a million occurrences.[7]

[7]The query refers to the social network site `vk.com`.

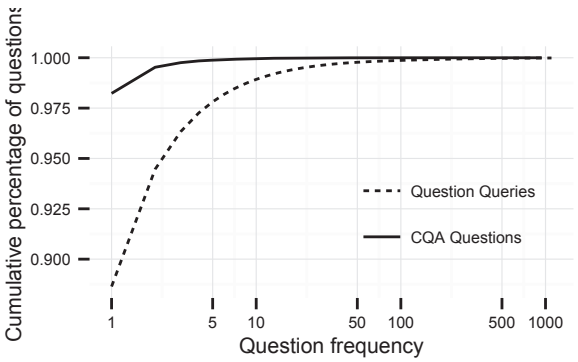

Figure 3: Question frequency in the CQA and question queries datasets.

Figure 4 shows the question length distributions in both datasets. While almost all the question queries have at most ten words, only one third of the CQA questions are that "short." (but note that we combine the question title and description fields). The average question query has a length of about six to seven words (about five to six not counting question words); the longest having 114 words probably copy-pasted from an exam and not reprinted here for space restrictions. The average CQA question is much longer with about 24 words (28 including question words) and the longest CQA question is about 1,000 words including its description.

4. QUESTION QUERY CLASSIFICATION

In order to infer a category assignment for the queries in the question query log, we employ a machine learning approach: Using the CQA questions and their assigned categories as a training set, we train a classifier that predicts the categories of unlabeled search engine question queries with high accuracy. To this end, we first derive different feature representations from the question queries and CQA questions, using the representation strategies outlined below. We compare a bag-of-words representation, which is effective but unwieldy due to its amount of features and only applicable to a subset of the question query data, to a much more compact topic-model-based representation. Since we finally want to classify the question queries, the general process of transferring the trained classifier from CQA data to question queries is as follows: First, we extract features (bag-of-words or topic models) from the question queries, then train a classifier with these features on the CQA questions, and finally apply the classifier to the question queries. The (unsupervised) feature extraction from the target dataset ensures better transferability of the classifier.

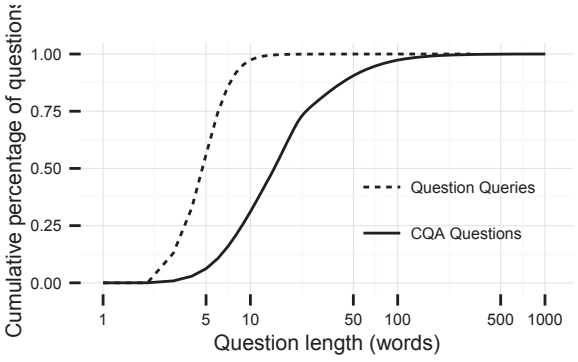

Figure 4: Question length in the CQA and question queries datasets.

4.1 Bag-of-words Features

Our first model is based on a bag-of-words representation, where each question is represented as a term frequency vector of (case-folded and lemmatized) unigrams. A bag-of-words model can be very complex: across the entire question query dataset, there are more than 16 million distinct words. Since we are using the CQA questions as our training set, our classifier can only consider the 1.3 million words that also occur in CQA questions. Out of this number, we retain only those words that occur in at least ten question queries, resulting in 137,032 features for our question query representation.

Besides its complexity, the main drawback of the bag-of-words model is the divergence of the feature sets between the two datasets. Out of our nearly one billion question queries, only 85% contain vocabulary from the bag-of-words model. We employ probabilistic topic modeling in order to reduce the model complexity, as well as to improve the transferability of the classifier. We investigate two different probabilistic topic models: Latent Dirichlet Allocation [8] and the Biterm Topic Model (BTM) proposed by Cheng et al. [13].

4.2 Topic Model Features

Both LDA and BTM are generative Bayesian models that uncover latent topics in a given text corpus by modeling the formation of documents as the result of a probabilistic process. For the purposes of feature derivation for our classification task, they operate in two basic steps, which can be summarized as *inference* and *representation*. The inference step involves finding the model parameters that best fit the observed data (the questions/documents in the corpus) for a given topic number k. Given a topic model thus trained, documents can be represented as k-vectors of topic probabilities. The generative model that is assumed to have generated the observed documents differs significantly between LDA and BTM.

From the LDA perspective, each word in each document is generated by first drawing a topic from a document-specific topic distribution, and then drawing the word from the word distribution for that topic. In order to accurately infer the per-document topic distributions, LDA depends on document-level context, and tends to perform poorly on short texts where word co-occurrence information is sparse [13].

The Biterm Topic Model circumvents the data sparseness problem by modeling term co-occurrence directly: In BTM's generative model, documents are modeled as sets of co-occurring words (biterms). Each biterm in a given document is generated by drawing a topic from a single global topic distribution, and then drawing the biterm from that topic's biterm distribution.

The benefit of representing documents as vectors of latent topic probabilities is two-fold—first, the representation is much more compact than a bag-of-words model of similar performance, and second, it captures high-level semantic

structure based on unigram occurrence alone, allowing a larger fraction of the question query log to be classified.

Our topic-model-based classification pipeline operates as follows: we apply stopword removal, case folding and lemmatization to all datasets. We then fit topic models to the question query dataset with the topic count k ranging from 10 to 500. For LDA, we employ the implementation available as part of the *gensim* software package.[8] To fit Biterm Topic Models, we use the implementation maintained by one of the BTM authors.[9] We then represent the CQA questions using the models fitted to the question queries and split the CQA questions into a training and validation set, comprising 70% and 30% of the questions, respectively. We use the validation set to select the best performing topic model, which we then evaluate on the web search question test set.

In the following section, we describe the results of our classification experiments and insights on questioning behavior in the search engine log.

5. EXPERIMENTAL RESULTS

According to the above procedure, we train a multinomial naïve Bayes classifier on the CQA training set for each of our question query models, and compare their performance on the CQA validation set. Having selected the best performing models from this run, we train new classifiers on the entire CQA data and evaluate them using the web search question test set. In order to compare the performance of the different models, we compute the classification performance on each target class, and then average over the classes to arrive at the macro-average precision and recall, as defined by [27]. Finally, we classify all questions in the web search log in order to gain further insights into what users ask search engines. For the classification experiments described below, we employ the multinomial naïve Bayes implementation from the Apache Spark MLlib library.[10]

5.1 Performance on CQA Questions

In order to compare the performance of the different topic models on the CQA data, we first fit a topic model to the question query data for the different numbers of topics. Due to the large amount of input data, this is a time consuming process; fitting the 500-topic BTM model requires approximately 80 hours of wall-clock time on a machine with sixteen 1.6 GHz CPU cores, while the largest LDA model requires about 24 hours. For both topic models, we use an incremental variant of the inference algorithm. Our observations confirm those of [13]— while the processing time for BTM is higher than for LDA, the memory requirements are lower.

Figure 5 shows the classification performance of the topic model-based features on the validation set, with the number of latent topics ranging from ten to 500. The biterm topic model outperforms LDA by a large margin for all topic counts. Considering the sparse word co-occurrence information found in web search queries, this result confirms our expectations. Both topic models' performance increases with growing number of topics, but the effect is more pronounced for LDA. More fine-grained latent topics make more informative features for query categorization in both cases. While the bag-of-words model outperforms both BTM and LDA,

[8] https://github.com/piskvorky/gensim

[9] https://github.com/xiaohuiyan/OnlineBTM

[10] https://spark.apache.org/mllib/

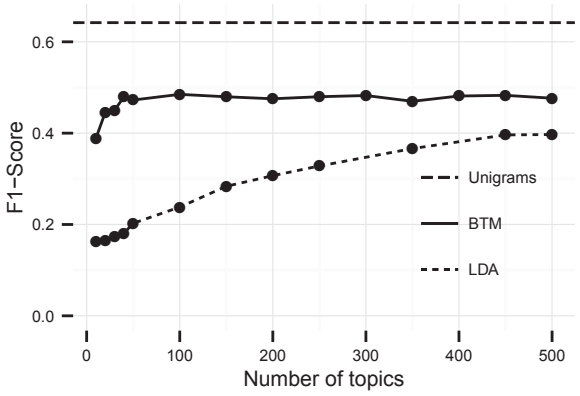

Figure 5: Classification performance of the topic model features on the CQA validation set.

it is at the cost of greater model complexity: the number of dimensions in the feature vector for the bag-of-words model is four orders of magnitude larger.

5.2 Performance on Web Search Test Data

Based on the above results, we conclude that the biterm topic model is better suited than LDA to our application domain. Hence, we compare the performance of the BTM features to the bag-of-words features on the web search query test set. As a simple baseline for comparison, we implement a majority-vote classifier based on CQA retrieval. To this end, we index the CQA dataset using the Okapi BM25 retrieval model, which has served as a baseline in previous studies on CQA retrieval [32]. At classification time, we submit the unlabeled query to this index, and pick the most common category among the ten first search results. In case of ties, we pick the category with the higher aggregate retrieval score.

The results of our comparison are summarized in Table 4, where we show the test set performance of CQA retrieval, the bag-of-words model, and the BTM models which perform best on the validation set. The rightmost column of the

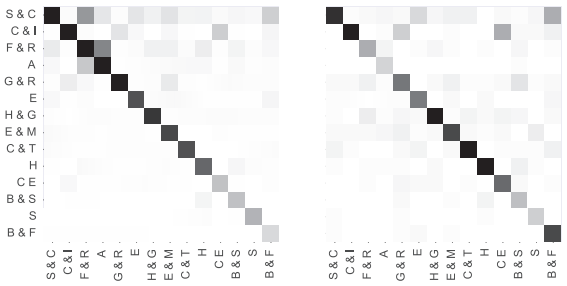

Figure 6: Confusion matrices for the bag-of-words classifier on the CQA validation set (left) and the question queriest test set (right). The rows are the true classes, the columns are the predictions; the ordering of the classes is the same as in Table 2.

Features	Precision	Recall	F_1-Score	Gain
CQA Retrieval Baseline				
—	0.67	0.66	0.66	—
Bag-of-words				
137,032	0.61	0.7	0.65	+2%
Biterm Topics				
100	0.47	0.53	0.50	+1%
200	0.46	0.49	0.47	±0%
300	0.46	0.50	0.48	±0%
400	0.46	0.50	0.48	±0%
450	0.49	0.53	0.51	+4%

Table 4: Performance of the Bag-of-words and BTM models on the web search query test data. The final column shows the change in F_1-score relative to the validation set.

table shows the relative performance gain (or loss) incurred in the transfer from the CQA to the web search data. While bag-of-words features still outperform the topic model on the test set, the difference in F1-score between bag-of-words and the best-performing BTM model is smaller compared to the validation set.

Being the best-performing of our machine learning models, we select the bag-of-words classifier to investigate the topic distribution in the web search question dataset; for the purpose of our post-hoc analysis, classification speed is not a major concern. However, in a live retrieval setting, we argue that one may prefer the BTM classifier despite its lower performance: due to the more compact feature vector, classification with BTM is much faster; on a 100-node Hadoop cluster running many classifications in parallel, the bag-of-words classifier requires on average three milliseconds of CPU time to classify a single question, compared to 1.3 milliseconds for the BTM classifier.

As shown in Figure 6, the classifier succeeds at distinguishing most of the categories rather well. Two exceptions are the "Family & Relationships" and "Adult" categories, which are frequently confused, as well as the "Computers & Internet" and "Consumer Electronics" categories. In both cases, a likely explanation is the natural overlap in vocabulary between these pairs of categories.

While the CQA retrieval classifier achieves a slightly higher F1-score than bag-of-words on the web search question test set, it incurs a much larger computational overhead—an average of 407 milliseconds per query, with the index stored on a solid-state disk. More advanced retrieval models have been shown to outperform BM25 in terms of CQA retrieval performance [32]. However, the overhead of an index lookup for each classification may prove prohibitive in a live retrieval setting.

5.3 Categorizing Web Search Questions

Below, we showcase some of the insights gained from the category distribution of the question queries in our query log. Since even our three human annotators were unable to favor a category assignment by majority in 17% of the cases, and our classifier agrees with them only two thirds of the time, the category of any individual query should be taken with a grain of salt. However, we do consider our model good enough to study general trends in the data.

Figure 7 shows the distribution of question query categories by month over the entire dataset. The shading in the cells

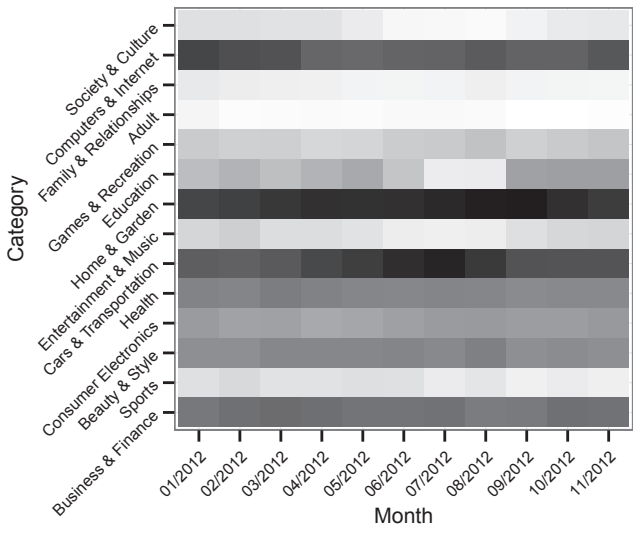

Figure 7: Distribution of monthly question query volume over categories. For each month, the shadings of the grid cells represent the categories' relative contributions to that month's total number of queries.

shows the contribution of each category to the total query volume for the corresponding month. The category axis is ordered by descending frequency in CQA questions, for easy comparison with Table 2. The category distribution our classifier infers for the web search questions is quite different from the distribution of category sizes among the CQA data. For instance, "Home & Garden" is the largest web search question category, covering over 13% of the web search queries, as opposed to 5% of CQA questions. Only 4% of web search queries are assigned to the "Society & Culture" category, compared to 18% of CQA questions.

Beyond this, the development of categories' query volume over time is of interest. While the query volume for some categories, such as "Health" or "Beauty & Style," remains more or less constant throughout the year, others show a pronounced seasonal variation. Most notably, the "Education" category reaches its low point during the months of July and August, while "Cars & Transportation" peaks around the same time. This may reflect askers embarking on their summer vacations, and abandoning education-related inquiries for travel-related ones.

Figure 8 shows the category distribution for some prominent question prefixes and suffixes. For the prefixes, we select a set of *how-to* question prefixes that are the most frequent in the query log, and compute the category proportions for the questions starting with each prefix. Some prefixes are strongly correlated with a single category, such as the `[how to cook ...]` questions with "Home & Garden." Other question prefixes, like `[how to make ...]` or `[how to learn ...]` are more evenly split among categories and occur to some extent in each one. As a side benefit, this analysis provides a sanity check for our classification model: expressions with several plausible contexts are distributed across the appropriate categories. For instance, the `[how to clean ...]` questions, with their corresponding housekeeping-, computer-

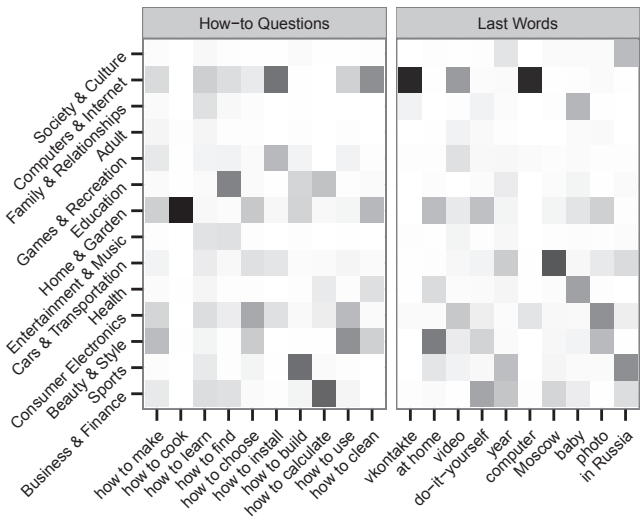

Figure 8: Distribution of common *how-to* questions and last words over categories. The items on the x-axis are ordered by total number of occurrences in the query log, highest at the left. For each item, the shadings of the grid cells represent the categories' relative contributions to that item's occurrences.

maintenance-, and personal-hygiene-related contexts, appear most frequently in the "Home & Garden," the "Computers & Internet," and the "Beauty & Style" categories, respectively.

In the right half of Figure 8, we show the distribution of the ten most common question endings over the categories, which reveals similar patterns to the questions' initial parts.

In an additional avenue of inquiry, we investigate the prevalence of advanced search operators—such as quoting, boolean expressions, or restricting the search to certain domains or file types—among question queries. Studies of general query logs have found a single-digit percentage of queries to use operators. For instance, [31] report 1.12% of queries recorded over a 13-week period containing operators, and 8.7% of users employing operators at least once during that time. We conjecture that among users who formulate queries as natural-language questions, operator use will be even rarer. Indeed, out of the billion question queries in our dataset, only 0.2% contain any search operators; only 1% of the 145 million unique users use operators at all.

In our query log, the quotation operator for phrasal search is by far the most prevalent, accounting for about 96% of all operator occurrences. Well-known operators like quotation and exact word match are equally prevalent across all categories, while the use of more advanced functionality often appears concentrated to a single category. For instance, the word distance operators (for retrieving only documents where the query terms occur within a user-specified distance) occur most often in the "Education" category.

6. CONCLUSION AND OUTLOOK

We have conducted the first large-scale analysis of non-English question querying behavior on a web search engine. Our main goal was to analyze the categories that searchers are interested in over the time of one year. To this end we have based our study on the about 1 billion questions submitted to a large commercial search engine in 2012.

As for the classification of the questions, we could not follow the practice used for classifying general web queries. There, established technologies use the search results to enrich the short query strings and to classify a query based on the results or the documents clicked by a user; however, in the case of questions that are rarely submitted by more than one user, click-through is much sparser. Since we also had in mind to develop a classifier that can be used in an online search engine, the fact that result information is not available for most of the questions ruled out the use of the standard procedure. Contrary to query classification for ad-matching or classification of questions at question answering platforms, that often classify into huge hierarchies with many classes, we aim at a flat set of a few categories only that can be easily integrated as additional features in the retrieval process (e.g., to select appropriate verticals).

Our suggested approach to question query classification is to use features extracted from the question queries to train a classifier on labeled CQA questions (where the asker assigns categories to posted questions) and then transfer this classifier back to the web search question queries. Our experiments show this approach to work very well given the 14 target classes. Hence, even though studies have shown that users tend to submit different questions to search engines than to CQA services, a fact also visible in our analyses, the classification transferability is not harmed. Training the classifiers on all the questions posted to a CQA service in the same year as the search engine questions, an F-measure of about 0.5 shows a decent performance given the 14 classes. Interestingly, the accuracy of the very efficient biterm topic model-based classifier is not much worse than the less efficient bag-of-words-based classifiers that had been proposed in previous studies for question classification.

Our experimental study of the year-long question query log shows some interesting first insights on categorized question asking behavior on a non-English search engine. Not too surprisingly, education questions are hardly observed in the months of summer vacation, while travel questions have their peak appearance in this time. The ratio of questions related to home and garden or health is rather stable over the year, while not too surprisingly "adult" topics are much less present in questions than in general web search queries. Further analyses on how-to questions, the questions' last words, and search operator use, also revealed some interesting insights.

Still, our first analyses should be seen as a starting point to use the question query classification for future work that can help improve retrieval performance on questions by better tailoring the results to the users' needs. This is especially important for questions that cannot directly be answered by showing related CQA questions. The amount of questions not directly answerable from CQA data is still an important direction for future research. Complementing our results with a similar study of question categories on English questions could shed some light on cultural differences in asking behavior and might help search engines to better address the different markets. Since the sheer amount of questions in the total query stream still is increasing, such topics will only get more important in the future. Potential applications abound—for instance in mobile voice search—to enable users to more naturally interact with retrieval systems via questions.

Acknowledgements

We thank Yandex (Alexey Gorodilov, Pavel Serdyukov, and Alexander Sadovskiy) and Mail.Ru (Andrey Oleynik) for preparing the datasets and granting access. The reported study was conceived during Pavel Braslavski's research stay at Bauhaus Universität Weimar supported by DAAD in fall 2013 and advanced during his visit supported by MUMIA network in October 2014. Pavel Braslasvki's and Galina Lezina's current work is partially supported by RFBR, research project #14-07-00589-a.

References

[1] Anne Aula, Rehan M. Khan, and Zhiwei Guan. How does search behavior change as search becomes more difficult? In *Proceedings of CHI 2010*, pages 35–44.

[2] Peter Bailey, Ryen W. White, Han Liu, and Giridhar Kumaran. Mining historic query trails to label long and rare search engine queries. *ACM Transactions on the Web*, 4(4): 15, 2010.

[3] Judit Bar-Ilan, Zheng Zhu, and Mark Levene. Topic-specific analysis of search queries. In *Proceedings of the WSCD 2009 Workshop*, pages 35–42.

[4] Steven M. Beitzel, Eric C. Jensen, Abdur Chowdhury, Ophir Frieder, and David Grossman. Temporal analysis of a very large topically categorized web query log. *Journal of the American Society for Information Science and Technology*, 58(2):166–178, 2007.

[5] Steven M. Beitzel, Eric C. Jensen, David D. Lewis, Abdur Chowdhury, and Ophir Frieder. Automatic classification of web queries using very large unlabeled query logs. *ACM Transactions on Information Systems*, 25(2):9, 2007.

[6] Jerome R Bellegarda. Spoken language understanding for natural interaction: The Siri experience. In *Natural Interaction with Robots, Knowbots and Smartphones*, pages 3–14. 2014.

[7] Michael Bendersky and W. Bruce Croft. Analysis of long queries in a large scale search log. In *Proceedings of the WSCD 2009 workshop*, pages 8–14.

[8] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[9] Andrei Z. Broder, Marcus Fontoura, Evgeniy Gabrilovich, Amruta Joshi, Vanja Josifovski, and Tong Zhang. Robust classification of rare queries using web knowledge. In *Proceedings of SIGIR 2007*, pages 231–238.

[10] Li Cai, Guangyou Zhou, Kang Liu, and Jun Zhao. Large-scale question classification in CQA by leveraging Wikipedia semantic knowledge. In *Proceedings of CIKM 2011*, pages 1321–1330.

[11] Xin Cao, Gao Cong, Bin Cui, and Christian S. Jensen. A generalized framework of exploring category information for question retrieval in community question answer archives. In *Proceedings of WWW 2010*, pages 201–210.

[12] Wen Chan, Weidong Yang, Jinhui Tang, Jintao Du, Xiangdong Zhou, and Wei Wang. Community question topic categorization via hierarchical kernelized classification. In *Proceedings of CIKM 2013*, pages 959–968.

[13] Xueqi Cheng, Yanyan Lan, Jiafeng Guo, and Xiaohui Yan. BTM: Topic modeling over short texts. *IEEE Transactions on Knowledge and Data Engineering*, paper 1, 2014.

[14] Huizhong Duan, Yunbo Cao, Chin-Yew Lin, and Yong Yu. Searching questions by identifying question topic and question focus. In *Proceedings of ACL 2008*, pages 156–164.

[15] Baichuan Li, Irwin King, and Michael R Lyu. Question routing in community question answering: Putting category in its place. In *Proceedings of CIKM 2011*, pages 2041–2044.

[16] Xiao Li, Ye-Yi Wang, and Alex Acero. Learning query intent from regularized click graphs. In *Proceedings of SIGIR 2008*, pages 339–346.

[17] Jian Liu, Yiqun Liu, Min Zhang, and Shaoping Ma. How do users grow up along with search engines?: A study of long-term users' behavior. In *Proceedings of CIKM 2013*, pages 1795–1800.

[18] Qiaoling Liu, Eugene Agichtein, Gideon Dror, Evgeniy Gabrilovich, Yoelle Maarek, Dan Pelleg, and Idan Szpektor. Predicting web searcher satisfaction with existing community-based answers. In *Proceedings of SIGIR 2011*, pages 415–424.

[19] Qiaoling Liu, Eugene Agichtein, Gideon Dror, Yoelle Maarek, and Idan Szpektor. When web search fails, searchers become askers: Understanding the transition. In *Proceedings of SIGIR 2012*, pages 801–810.

[20] Meredith Ringel Morris, Jaime Teevan, and Katrina Panovich. What do people ask their social networks, and why?: A survey study of status message Q&A behavior. In *Proceedings of CHI 2010*, pages 1739–1748.

[21] Bo Pang and Ravi Kumar. Search in the lost sense of query: Question formulation in web search queries and its temporal changes. In *Proceedings of ACL 2011*, pages 135–140.

[22] Greg Pass, Abdur Chowdhury, and Cayley Torgeson. A picture of search. In *Proceedings of Infoscale 2006*, paper 1.

[23] Bo Qu, Gao Cong, Cuiping Li, Aixin Sun, and Hong Chen. An evaluation of classification models for question topic categorization. *Journal of the American Society for Information Science and Technology*, 63(5):889–903, 2012.

[24] Matthew Richardson. Learning about the world through long-term query logs. *ACM Transactions on the Web*, 2(4): 21, 2008.

[25] Johan Schalkwyk, Doug Beeferman, Françoise Beaufays, Bill Byrne, Ciprian Chelba, Mike Cohen, Maryam Kamvar, and Brian Strope. "Your word is my command": Google search by voice: A case study. In *Advances in Speech Recognition*, pages 61–90. 2010.

[26] Dou Shen, Jian-Tao Sun, Qiang Yang, and Zheng Chen. Building bridges for web query classification. In *Proceedings of SIGIR 2006*, pages 131–138.

[27] Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437, 2009.

[28] Amanda Spink and H. Cenk Ozmultu. Characteristics of question format web queries: An exploratory study. *Information processing & management*, 38(4):453–471, 2002.

[29] Amanda Spink, Bernard J. Jansen, Dietmar Wolfram, and Tefko Saracevic. From e-sex to e-commerce: Web search changes. *Computer*, 35(3):107–109, 2002.

[30] Ingmar Weber, Antti Ukkonen, and Aris Gionis. Answers, not links: Extracting tips from Yahoo! Answers to address how-to web queries. In *Proceedings of WSDM 2012*, pages 613–622.

[31] Ryen W. White and Dan Morris. Investigating the querying and browsing behavior of advanced search engine users. In *Proceedings of SIGIR 2007*, pages 255–262.

[32] Xiaobing Xue, Jiwoon Jeon, and W. Bruce Croft. Retrieval models for question and answer archives. In *Proceedings of SIGIR 2008*, pages 475–482.

[33] Zhe Zhao and Qiaozhu Mei. Questions about questions: An empirical analysis of information needs on Twitter. In *Proceedings of WWW 2013*, pages 1545–1556.