Towards Query Logs for Privacy Studies: On Deriving Search Queries from Questions

Asia J. Biega[1](✉), Jana Schmidt[2], and Rishiraj Saha Roy[2]

[1] Microsoft Research, Montreal, Canada
asia.biega@microsoft.com
[2] Max Planck Institute for Informatics,
Saarland Informatics Campus, Saarbrücken, Germany
{jschmidt,rishiraj}@mpi-inf.mpg.de

Abstract. Detailed query histories often contain a precise picture of a person's life, including sensitive and personally identifiable information. As sanitization of such logs is an unsolved research problem, commercial Web search engines that possess large datasets of this kind at their disposal refrain from disseminating them to the wider research community. Ironically, studies examining privacy in search often require detailed search logs with user profiles. This paper builds on an observation that information needs are also expressed in the form of questions in online Community Question Answering (CQA) communities. We take a step towards understanding the process of formulating queries from questions to form a basis for automatic derivation of search logs from CQA forums. Specifically, we sample natural language (NL) questions spanning diverse themes from the StackExchange platform, and conduct a large-scale conversion experiment where crowdworkers submit search queries they would use when looking for equivalent information. We also release a dataset of 7,000 question-query pairs from our study.

1 Introduction

Background. Commercial Web search engines refrain from disseminating detailed user search histories, as they may contain sensitive and personally identifiable information[1]. Studies examining privacy in search, however, require extensive search logs with user profiles to examine the sensitive semantics of queries or the topical distribution of user interests [1,5,6,8,16].

While there exist a number of public search query logs, none of them contain *detailed user histories*. Relevant among these, the TREC Sessions Track 2014 data [7] has 148 users, 4.5k queries, and about 17k relevance judgments. There are roughly ten sessions per user, where each session is usually a set of reformulations. Such collections with just a couple of queries per user are inadequate

[1] https://en.wikipedia.org/wiki/AOL_search_data_leak.

J. M. Jose et al. (Eds.): ECIR 2020, LNCS 12036, pp. 110–117, 2020.
https://doi.org/10.1007/978-3-030-45442-5_14

for driving research in privacy, especially research that focuses on topical profiling. The 2014 Yandex collection [15] is useful for evaluating personalization algorithms. However, to protect the privacy of Yandex users, every query term is replaced by a numeric ID. This anonymization strategy makes semantic interpretation impossible and may be a reason why this collection has not received widespread adoption in privacy studies. Interpretability of log contents is vital for understanding privacy threats [5,6,8].

Motivated by the lack of public query logs with rich user profiles, Biega et al. [5] synthesized a query log from the StackExchange platform[2] – a collection of CQA subforums on a multitude of topics. Queries in the synthetic log were derived from users' information needs posed as natural language questions. A collection like this has three advantages. First, it enables creation of rich user profiles by stitching queries derived from questions asked by the same user across different topical forums. Second, since it is derived from explicitly public resources created by users under the StackExchange terms of service (allowing reuse of data for research purposes), it escapes the ethical pitfalls intrinsic to dissemination of private user data. Third, CQA forums contain questions and assessments of relevance in the form of accepted answers *from the same user*, which is vital for the correct interpretation of query intent [2,9]. Other signals like similar queries and reformulations can also be simulated with related questions and duplicates, available on most CQA forums.

Contributions. We take a step towards better automatic question-query derivation methods to improve on the approach taken by Biega et al. [5] where queries are constructed by choosing a random number of terms with highest TF-IDF scores. An accurate approach like this would enable the creation of high-quality search collections down the road. We make the following contributions: (1) We conduct a large-scale user study where crowdworkers convert questions to queries, controlling for several biases; (2) We provide insights from the collected data that could drive strategies for automatic conversion at scale and be used to derive synthetic search collections for privacy studies; (3) We release 7,000 question-query pairs collected from the study[3].

2 Setting up the User Study

Filtering Subforums. We used the StackExchange dump[4] from March 2018 with data for more than 150 different subforums. We are interested in textual questions in English and thus exclude forums primarily dealing with programming, mathematics, and other languages. Moreover, we want to avoid highly-specialized forums as an average AMT user may not have the background knowledge to generate queries for niche domains. We thus excluded all subforums with

[2] https://stackexchange.com/sites.

[3] https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/impact/mediator-accounts/.

[4] https://archive.org/details/stackexchange-snapshot-2018-03-14.

less than 100 questions, as a proxy for expression of a critical mass of interest, leaving us with a total of 75 subforums.

Sampling Questions. As a proxy for questions being understandable by users, we choose only those that have an answer accepted by the question author, and with at least five other answers provided. Under this constraint, we first sample 50 subforums from the 75 acceptable ones to have high diversity in question topics. Next, we draw 100 questions from each of these subforums, producing a sample of 5000 questions to be used as an input in the main study.

Setup. We recruited a total of 100 AMT Master workers[5] who had an approval rate of over 95%, to ensure quality of annotations. A unit task, i.e., an AMT HIT (Human Intelligence Task) consisted of converting *fifty NL questions* to Web queries to capture *user-specific* querying traits (thirty in our pilot study). Since this is significant effort expected to require more than an hour's work at a stretch, we paid $9 per HIT ($6 in our pilot, owing to fewer questions). The workers (Turkers) were given three hours to complete a HIT, while the actual average time taken turned out to be 1.6 h. This is about two minutes per question, which we deem as a reasonable time required for understanding the intent of a typical CQA question that often has a few hundred words.

Guidelines. Guidelines were kept to a minimum to avoid biasing participants towards certain query formulation behavior: they only stated the requirement of building a search query aimed at retrieving equivalent information as the source question. We provided five examples to better illustrate the task, that were meant to cover the various ways of arriving at a reasonable query. To build queries, we allowed workers to: (i) select exact words from the text of the question, (ii) modify question words (*'use'* ↦ *'using'*), or, (iii) use their own words to clarify the information need. These cases were not made explicit, but communicated by coloring words in the text and the query. Questions were presented as follows (some choices aimed at avoiding title bias, see Sect. 3.2):

$$[Subforum\ name]\ \textit{Title Body}$$

Each question was a concatenation of the StackExchange post title and its body, prefixed with the subforum name of the post for context. The main task was accompanied by a *demographic survey* to help us understand if such features influence how people formulate queries.

Pilot Study. We tested the setup with a pilot containing five HITs with 30 questions each. The average query length came out to be 5.7 words with a standard deviation of 2.4 words. Out of the 150 questions in total, the forum name was included in the corresponding query 33 times. In nine of such cases, the subforum name was not present in the title or body of the question, which suggests that the presence of the subforum name is important in disambiguating the context. Most query words were chosen from the title, although title words are often repeated in the body of the question. Workers used their own words

[5] https://www.mturk.com/help#what_are_masters.

or words modified from the question 47 times. These results suggest that participants generally understood the instructions, and gave us the confidence that this setup can be used in the main study.

3 Conducting the User Study

Data Collection. In the main study, we asked 100 AMT users to convert 5000 questions to queries (50/Turker). Users who participated in control studies were not allowed to take part again, to *avoid familiarity biases* arising from such repetition. Guidelines were kept the same as in the pilot study. The mean query length was now 6.2 words: this reflects high complexity in the underlying information needs, and in turn, interesting research challenges for methods aiming at automated conversion strategies for query log derivation. Key features of the final dataset include: (i) question topics spanning 50 different subforums of StackExchange, and (ii) question-query pairs grouped by annotator IDs, making the testbed suitable for analyzing user-specific query formulation.

3.1 Analysis

We looked into three aspects of *question-query pairs* when trying to discriminate between words that are *selected* for querying, and those that are not.

Position. We measured relative positions of query and non-query words in the question, and found that a major chunk ($\simeq$60%) of the query words originate from the first 10% of the question. The next 10% of the question contributes an additional 17% of words to the query; the remaining 80% of the question, in a gently diminishing manner, produce the remaining 13% of the query. This is a typical top-heavy distribution, suggesting humans conceptualize the core *content* of the information need first and gradually add specifications or conditions of *intent* [13,14] towards the end. Notably, even the last 10% of the question contains 2.78% of the query, suggesting that we cannot disregard tail ends of questions. Finally, note that the title is positioned at the beginning of the question (Sect. 3.2), and alone accounts for 57% of the query. Title words, however, do repeat in the body. Further inspection reveals that only 12% of the query mass is comprised of words that appear exclusively in the title, signifying importance of the body. We also allowed users to use their *own words* in the queries. Our analysis reveals that a substantial 17% of query words fell into this category. Such aspects of this data pose interesting challenges for query generative models.

Part-of-Speech (POS). Words play various roles in NL, with a high-level distinction between *content words* (carrying the *core* information in a sentence) and *function words* (specifying *relationships* between content words). Web users have a mental model of what current search engines can handle: most people tend to drop function words (prepositions, conjunctions, etc.) when issuing queries [4], perhaps believing those are of little importance in query effectiveness. These intuitions are substantiated by our measurements: content words (nouns, verbs, adjectives, and adverbs) account for a total of 79% (47%, 15%, 13%, and 4%, respectively) of the query, while function words constitute only 21% of the query. For interpretability, we use the 12 Universal POS tags (UTS)[6]. Our findings partially concur with POS analysis of Yahoo! search queries from a decade back [4] where nouns and adjectives were observed to be the two most dominant tags; verbs featured in the seventh position with 2.4%. We believe that differences in our scenario can be attributed to more complex information needs that demand more content words to be present in queries. These insights from the POS analysis of queries can be applied to several tasks, like query segmentation [10].

Frequency. A verbose information need may be characterized by certain recurring units, which prompted us to measure the normalized frequency TF_{norm} of a term t in a question Q, as $TF_{norm}(t, Q) = TF(t, Q)/len(Q)$, where $len(Q)$ is the question length in words. Query terms were found to have a mean TF_{norm} of 0.032, significantly higher than that of non-query terms (0.018).

3.2 Control Studies

Title Position Bias. A vital component of any crowdsourced study is to check if participants are looking for quick workarounds for assigned tasks that would make it hard for requesters to reject payments, and to control for confounding biases. In the current study, a major source of bias stems from the fact that a question is not just a sequence of words but a semi-structured concept (subforum, title, body.) Web users might be aware that question titles often summarize questions. Thus, if the structure is apparent to the annotator, they might use words only from titles without examining the full question content.

To mitigate this concern, we present titles in the same font as the body, and do not separate them with newlines. Nevertheless, users may still be able to figure out that the first sentence is indeed the question title. To quantify such *position bias* of the title, we used ten HITs (500 questions) as a *control experiment* where, unknown to the Turkers, the title was appended as the *last sentence* in the question. These 500 questions were also annotated in the usual setup in the main study. We compare the main and the control studies by measuring how often users chose words from the first and the last sentences (Table 1). Values were normalized by the length of the question title, as raw counts could mislead the analysis (longer question titles contribute larger *numbers* of words to queries).

We make the following observations: (i) in both the main study and the control, users choose words from titles very often (≃97% and ≃84%, respectively), showing similar task interpretation. Note that such high percentages are acceptable, as question titles typically do try to summarize intent. (ii) Relatively similar percentages of query words originate from titles in both cases (37.7% vs. 26.1%). (iii) If Turkers were trying to do the task just after skimming the first sentence (which they would perceive as the title), the percentage of words from

[6] https://github.com/slavpetrov/universal-pos-tags.

the first sentence in the control would have been far higher than a paltry 12.2%, and the last sentence would contribute much lower than 26.1%. We also observed that in 4.1% of the cases, words were chosen *exclusively* from the last sentence.

Table 1. Measurements from the position bias control study.

Property	Main study	Control study
Times question title word chosen for query	96.6%	83.8%
Question title words in query	37.7%	26.1%
First sentence words in query	37.7%	12.2%
Last sentence words in query	9.0%	26.1%

User Agreement. While the main focus of the study was to construct a sizable collection of question-query pairs, we were also interested in observing the effect of individual differences on query formulation. To this end, we issued ten HITs (each with fifty questions) completed by three workers each. The validity of the comparison comes from the experimental design where query construction is conditioned on a specific information need. We computed the average Jaccard similarity coefficient between all pairs of queries (q_1, q_2) for the same question: $J(q_1, q_2) = \frac{|q_1 \cap q_2|}{|q_1 \cup q_2|}$, where q_1 and q_2 are the sets of words of the compared queries. We find the average overlap to be 0.33; the overlap was observed to typically arise from the most informative question words, again indicating generally correct task interpretation. Such query *variability* has been explored in [3].

3.3 Crowdworker Demographics

We asked about crowdworkers' gender, age, country of origin, highest educational degree earned, profession, income, and the frequency of using search engines in terms of the number of Web queries issued per day (such activity could be correlated with "search expertise", and this expertise may manifest itself subtly in the style of the generated queries). From the 100 subjects in our study, coincidentally, female and male participation was exactly $50:50$. Nearly all workers lived in the USA except for three who lived in India. We found a weak positive correlation between the query length and age, and found that men formed slightly longer queries on average (6.56 words, versus 6.15 for women).

3.4 Dataset and Extended Analyses

The annotated dataset (with fields: study type, anonymous crowdworker ID, StackExchange user and post IDs, subforum name, post title, post body, and query) and an extended version of this paper with more analyses and details are available online[7]. The dataset contains $7,000$ (question, query) pairs in total: $5,000$ from the main study, 500 from the control experiment on title position bias, and $1,500$ from the control on user agreement.

4 Conclusions

We conducted a user study to provide a better understanding of how humans formulate queries from information needs described by verbose questions, and released $7k$ crowdsourced question-query pairs from 50 domains. Gaining insights into this process forms an important foundation for automated conversion methods to create rich public search collections useful in privacy studies of profiling and beyond. In addition to such algorithmic conversion, potential future directions include an analysis of the quality of crowdsourced queries [12] for our setup (such as their potential for retrieval), as well as applying our general methodology to other CQA datasets [11].

References

1. Adar, E.: User 4xxxxx9: anonymizing query logs. In: Proceedings of Query Log Analysis Workshop, International Conference on World Wide Web (2007)
2. Bailey, P., Craswell, N., Soboroff, I., Thomas, P., de Vries, A.P., Yilmaz, E.: Relevance assessment: are judges exchangeable and does it matter. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 667–674. ACM (2008)
3. Bailey, P., Moffat, A., Scholer, F., Thomas, P.: UQV100: a test collection with query variability. In: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 725–728. ACM (2016)
4. Barr, C., Jones, R., Regelson, M.: The linguistic structure of English web-search queries. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1021–1030. Association for Computational Linguistics (2008)
5. Biega, A.J., Saha Roy, R., Weikum, G.: Privacy through solidarity: a user-utility-preserving framework to counter profiling. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 675–684. ACM (2017)
6. Biega, J.A., Gummadi, K.P., Mele, I., Milchevski, D., Tryfonopoulos, C., Weikum, G.: R-susceptibility: an IR-centric approach to assessing privacy risks for users in online communities. In: Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, pp. 365–374. ACM (2016)
7. Carterette, B., Kanoulas, E., Hall, M., Clough, P.: Overview of the TREC 2014 session track. Technical report, Delaware University Newark (2014)
8. Chen, G., Bai, H., Shou, L., Chen, K., Gao, Y.: UPS: efficient privacy protection in personalized web search. In: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 615–624. ACM (2011)

[7] https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/impact/mediator-accounts/.

9. Chouldechova, A., Mease, D.: Differences in search engine evaluations between query owners and non-owners. In: Proceedings of the sixth ACM International Conference on Web Search and Data Mining, pp. 103–112. ACM (2013)
10. Hagen, M., Potthast, M., Beyer, A., Stein, B.: Towards optimum query segmentation: in doubt without. In: Proceedings of the 21st ACM International Conference on Information and Knowledge Management, pp. 1015–1024. ACM (2012)
11. Hagen, M., Wägner, D., Stein, B.: A corpus of realistic known-item topics with associated web pages in the ClueWeb09. In: Hanbury, A., Kazai, G., Rauber, A., Fuhr, N. (eds.) ECIR 2015. LNCS, vol. 9022, pp. 513–525. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-16354-3_57
12. Hauff, C., Hagen, M., Beyer, A., Stein, B.: Towards realistic known-item topics for the ClueWeb. In: Proceedings of the 4th Information Interaction in Context Symposium, pp. 274–277. ACM (2012)
13. Saha Roy, R., Katare, R., Ganguly, N., Laxman, S., Choudhury, M.: Discovering and understanding word level user intent in web search queries. J. Web Semant. **30**, 22–38 (2015)
14. Saha Roy, R., Suresh, A., Ganguly, N., Choudhury, M.: Place value: word position shifts vital to search dynamics. In: Proceedings of the 22nd International Conference on World Wide Web, pp. 153–154. ACM (2013)
15. Serdyukov, P., Dupret, G., Craswell, N.: Log-based personalization: the 4th web search click data (WSCD) workshop. In: Proceedings of the 7th ACM International Conference on Web Search and Data Mining, pp. 685–686. ACM (2014)
16. Zhang, S., Yang, G.H., Singh, L., Xiong, L.: Safelog: supporting web search and mining by differentially-private query logs. In: 2016 AAAI Fall Symposium Series (2016)