Building a Query Log via Crowdsourcing

Omar Alonso
Microsoft Corp.
Mountain View, CA
omalonso@microsoft.com

Maria Stone*
Yahoo! Inc.
Sunnyvale, CA
mariast@yahoo-inc.com

ABSTRACT

A query log is a key asset in a commercial search engine. Everyday millions of users rely on search engines to find information on the Web by entering a few keywords on a simple search interface. Those queries represent a subset of user behavioral data which is used to mine and discover search patterns for improving the overall end user experience. While queries are very useful, it is not always possible to capture precisely what the user was looking for when the intent is not that clear. We explore a different alternative based on human computation to gather a bit more information from users and show the type of query log that would be possible to construct.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Search process; J.4 [**Computer Applications**]: Social and Behavioral Sciences.

General Terms

Design, experimentation

Keywords

Crowdsourcing, query logs, query annotation, user studies

1. INTRODUCTION

The concept of information need is widely used in information retrieval to explain when a user requires any information to complete a task independent of the method used. A query is a way to express a part of such information need to a search system. In the case of a web search engine, a query log represents the information need from millions of users. Query logs are extremely useful because they contain user queries, IP addresses, timestamps, and other click data that are used to study, analyze, and improve many search engine features. Query logs are proprietary and such data is not available for external consumption. There are a few data sets that have been made available for research purposes and they only represent a small sample of the real traffic.

Search queries are entered by humans so it would be possible, in principle, to explore the idea of asking humans to share a query that they have performed recently. At the same time, we would like to take advantage of the request and ask them to provide a brief explanation of what they were looking for. This open call for sharing data can be implemented using crowdsourcing techniques.

What do we mean by crowdsourcing a query log? We would like to gather not only real queries but also the user's information need in the form of an explanation. Part of our work involves collecting some extra information that a true web query log lacks: an explicit description of the query intent by the same user who is issuing the query. Our goal is to construct a data set of $< q, a >$ pairs where q represents a query and a the annotation of such query by the user.

The research questions that we would like to ask are: It is possible to collect real queries and information needs from users via crowdsourcing? What are the characteristics of such query log? Which types of queries would users like to share? Would this data be useful? In this paper we explore the potential of our approach and present initial results.

There is previous research on gathering user data and annotating user information needs. The Lemur community query log project is an example of asking people to donate their queries [4]. The goal of the project was to create a database of web search activity by installing a toolbar in a browser. From the emerging topic of "games with a purpose", the human computation game Intentions aims to collect data about the intent behind search queries [3]. Zuccon *et al.* use crowdsourcing to gather search and interaction log data for interactive information retrieval [6]. Our approach uses a more direct path by using *crowdsourcing in the wild*. This is, a simple experiment that is executed in many platforms that asks users to provide a search pair that consists of a query and the corresponding information need.

If we look back at the early days of information retrieval, human computation played a pivotal role in the Cranfield research project for collecting documents, queries, and relevance assessments [1]. We can argue that such methodology included crowdsourcing for requesting *search questions* and relevance assessments using an expert crowd. Those search questions, provided by each expert, were "the reason for the research being undertaken leading to the paper". The letters

*Work done while author was affiliated with Microsoft.

SIGIR'14, July 6–11, 2014, Gold Coast, Queensland, Australia.

http://dx.doi.org/10.1145/2600428.2609479.

that were sent to each expert contain detailed instructions of what was expected [2].

We are interested in following a similar methodology for gathering pairs of queries and annotations in the spirit of the "reasons" behind search questions. We are aware of a number of differences. In contrast to Cranfield, we do not rely on an expert crowd and do not provide documents or web pages so users can provide queries. Our instructions are very simple and do not require a lot of work from our users.

In the rest of this paper, we describe the experimental design, data collection, analysis, and findings of our crowdsourced query log.

2. SURVEY DESIGN

Our survey design consists of asking workers to provide first what they were looking for and then, the exact query using the Human Intelligence Task (HIT) presented in Figure 1. Both questions were mandatory.

1. Please enter your search query for this search, exactly the way you entered it in the search box of your favorite search engine:
[Input text]
2. Please describe your search need in one or two sentences. What were you looking for?
[Input text]

Figure 1: HIT design template.

We tested two different version of the design: one with examples of correct and incorrect answers (E-examples) and a second one with no examples at all (E-noexamples). Table 1 shows the examples that were included as instructions in E-examples.

We conducted two separate batches of experiments B1 and B2 at different points in time, 2013 and 2014 respectively. The payment for each HIT was $0.05 and there was no prequalification required for any of the tasks.

On B1, we created the same task using design E-examples in Amazon Mechanical Turk (AMT) and CrowdFlower, two very popular platforms. We collected two data sets D1 and D2 that we describe as follows. After some basic quality control to remove spammers, D1 consists of 506 queries (471 unique) collected via AMT using the authors' personal accounts. D2 consists on 100 queries (98 unique) collected via CrowdFlower with multiple channels, also using one of the author's personal account. We report that in D2, we enforced more quality control and data cleaning than in D1.

Q1 acceptable	laptops under $500, "prarie dogs" habitat information, Napoleon death
Q1 unacceptable	http://www.google.com, bing.com, yahoo.com,
Q2 acceptable	Looking to buy a laptop for under $500, Doing background research about prarie dogs for my school report, Trying to find out when and how Napoleon died
Q2 unacceptable	Research, buy stuff, images

Table 1: Examples of acceptable and unacceptable queries and explanations.

Characteristic	D1	D2	D3	D4
Average terms in query	4.00	4.18	3.38	2.91
Stdev terms in query	2.26	2.52	2.1	2.31
Average terms in annotation	11.76	11.21	9.91	10.15
Stdev terms in annotation	5.76	6.05	5.62	7.26
1 term in query	9%	13%	16%	27%
2 terms in query	15%	12%	24%	23%
3 terms in query	23%	20%	22%	21%
> 3 terms in query	54%	55%	38%	28%

Table 2: Comparative statistics for the four data sets.

For B2, we created two tasks using designs E-examples and E-noexamples respectively in CrowdFlower only. We collected two more data sets D3 and D4 that we describe as follows. D3 consists on 256 queries (235 unique) and D4 consists on 332 queries (296 unique). We decided to rely more on CrowdFlower because they provide more channels for distributing the work and other quality control features such as contributor satisfaction. That is, what workers think about our task based on instructions, ease of job, and pay. We score an average of 4.3/5 in all tasks.

3. DATA ANALYSIS AND RESULTS

In Table 2, we report basic statistics about all data sets. The numbers from D1 and D2 are similar but they differ from what the literature reports on query log studies [5]. For example, average terms in queries is much higher than the reported ones (usually around 2.6 or 2.7). This is about a 65% increase in query length. The average terms in the annotation is around 10 words which is within bounds of the average English sentence length.

If we look at the results of query classification (done by the authors) in Table 3, informational queries tend to dominate the query set. The numbers for D1 and D2 are very similar. While D1 and D2 have the same amount of navigational queries, adult content was more present on D2. In terms of transactional queries, D2 has no such examples. D3 has a bit less in navigational and more on informational but comparable to the other two. We observe that the numbers are very different from the reported breakdown in the literature.

Recall that D1, D2, and D3 use design E-examples which contains examples of acceptable answers while D4 use E-noexamples with no examples. If we examine D4, the numbers are different from the other three data sets, in particular the increase of navigational queries. One possible explanation is that the use of examples as part of the instructions sets a limit on the type of queries that users were willing to share. Another example is the number of e-commerce

Query class	D1	D2	D3	D4
Navigational	8%	8%	6%	12%
Informational	91%	92%	94%	87%
Transactional	1%	0%	0%	1%

Table 3: Query classification for the four datasets.

Query	Annotation
Non-resident US Citizen IRA contributions	Doing research on if a non-US resident (but US citizen) can contribute to an IRA account.
What is the highest dose of metoprolol?	I was trying to find out what the highest dose was for the hypertension drug metoprolol.
samsung dvd burner	I was looking for an external dvd burner made by samsung that is portable
Wood drying rack	An outdoor collapsible drying rack to get price estimate
holly madison car accident	a picture holly madison posted after a car accident
titanic goofs	Factual scripting errors in the 1997 movie Titanic.
Eczema treatments	Over the counter treatments for Eczema
Sony A7R vs Olympus EM-10	mirrorless camera comparison
creation myth	I was looking for some myths about creation of the world like Genesis or Enuma Elish
ellen page	I was looking for Ellen Page's coming out speech at the HRC's Time to Thrive conference.
bitcoin news	I want to know what has happened with the Mt. Gox shutdown, and any other breaking news regarding bitcoin

Table 5: Sample of queries and information needs collected for informational queries.

Query	Annotation
instagc	instagc website
Instagc	Looking for link to website
instagc	gpt site
probux	to get to the probux website.
probux	This website
probux	the best PTC site web
facebook	I was trying to login my account
facebook	looking for signup in facebook
facebook	I was trying to get to my facebook
Facebook	The Facebook login

Table 4: Examples of information needs and queries collected for navigational queries.

queries that accounts for 5% in D3 and only 1% in D4. In general, we can say that D4 represents a more real data set.

We now describe the queries and annotations in more detail by showing some unedited examples (data from users is presented verbatim). Regarding navigational queries, users are very clear about what they want with explanations that are short and precise. Table 4 shows examples of queries and annotations. We can see that there is consensus regarding their respective information needs: go to a particular site.

In Table 5 we present a sample of queries categorized as informational. It is interesting to read the annotations and see how the query may not be capturing exactly what the user had in mind. The annotations provide more context about the query.

Exploring beyond informational, navigational, and transactional queries, we noticed clusters around seasonal content, current events, and how-to/task completion. Table 6 shows a sample of queries for seasonal topics in the United States such as Valentine's day and tax return. The Winter Olympics is also a common topic in D4 (gathered around February 2014). In Table 7 we present examples of queries where the user has a problem and looks for answers in a search engine. Some of those tasks are immediate (e.g., `facebook certificate expiration IE`) while others look more long term need (e.g., `bitcoin mining`).

Finally, how can we tell if the queries submitted are real? One possibility is to take the entire data set and see if those queries are available in a commercial log with the disadvantage that requires access to such infrastructure. Another solution would be to manually check against the auto-complete pull down menu and inspect if the completion matches the query (e.g., `tax -> tax forms`) but requires manual intervention. It is possible that users will submit made up queries but, at the same time, they are willing to share, in some cases, information that can be consider sensitive. For example, one user provided a number with the annotation `tracking number for UPS`. We entered the number on the search engine and it was indeed a valid UPS number.

4. CHALLENGES AND OPPORTUNITIES

Like any crowdsourcing-based experiment, task design and instructions determine the overall quality of the results. We tested a couple of designs and noted a difference in the outcome by eliminating the examples so users would not need to submit "right" queries. It is still unclear what is the best experiment to collect queries. We avoided imposing any sophisticated quality control mechanism for data collection with the exception of spam detection. Our original goal was to capture as much "raw" data as possible and we can report that it is possible.

How scalable is this approach? As long as people keep using search engines it should be possible to capture a sample of that traffic with human computation. That said, there is cost involved for sourcing the data as well as building quality control mechanism for acquiring the best data possible. Are workers providing real queries? We looked at all the queries and some of them do have a temporal pattern that corresponds to a current event. This could be a useful feature to assess how real the query set is. Would users share personal queries? We see some examples of adult queries and more personal content but probably less compared to a real query log. Privacy is an area that needs more exploration. So far, we asked workers for one single query, not a search session, and let them share the most representative query.

We observed that the queries collected from the experiments are by nature informational and we can argue that

Query	Annotation
sochi athletes	looking for the images of sochi athletes
Sochi men's skating results	Final results for Olympics men's figure skating short program
Sochi Olympics	Latest updates on Sochi Olympics
winter Olympics At Sochi us gold wins	progress of united states regarding gold medals
Valentine's day gift for men	I was just looking for some nice gift ideas for Valentine's day.
valentine's day history	I was looking for the year in which Valentine's Day became an official holiday.
Valentines Day Orginiality	I wanted to know how Valentines Day started.
What to do for Valentine's Day	Creative ways for valentine's day
tax forms	I wanted printable forms to work on my federal taxes.
tax forms	for particular tax forms to fill out my taxes

Table 6: Sample of queries grouped based on seasonality for the month of February 2014.

Query	Annotation
korg M1 battery replacement	Instructions on how to replace the internal battery on a Korg M1 keyboard (synthesizer). I wanted to see if it was something I could do myself or who to take it to and how much it would cost.
bitcoin mining	I wanted to learn how to use the mining pools and stratu proxies for mining bitcoins. Also the other different alt coins
Write Protected USB Pendrive	Trying to find out how to remove write protection from my USB pendrive
PTC strategy	I am new to PTC sites, but interested to know how to make some extra money, and wanted to see tips and ways to improve my money making ways.
boy scout patch uniform placement	I needed to know where to put the "trained" patch on the sleeve and if it was right or left.
facebook certificate expiration IE	I could not log into Facebook and kept getting a certificate expiration message. I was trying to figure out how to resolve it so I could get back on Facebook.

Table 7: Sample of queries grouped based by "how two" tasks.

they may represent the so called tail of the distribution. Further research needs to be done to better understand why users share certain type of queries. Perhaps navigational queries are seen as mere shortcuts than a query for a more concrete information need? Why users in D2 felt more confident sharing adult queries than those in D1?

We view the construction of crowdsourced query logs as an opportunity to create specific data sets that have a target in mind. For example, a set of temporal and seasonal queries, task-completion, and how-to queries. If we mine the annotations there is potential to identify a set of queries that users may find difficult to satisfy with current systems.

5. CONCLUSIONS AND FUTURE WORK

This initial work shows that constructing a crowdsourced query log should be feasible but not straightforward. We spent, on average, around $3 per 50 queries (including transaction costs) that indicates that it is possible to gather a reasonable data set using a modest budget fairly quickly. Users are willing to share their queries and tell a bit more by proving a concrete information need in the form of an annotation.

While our initial findings are encouraging, the size of the data collected is orders of magnitude inferior to a industrial query log. A number of open questions about the experimental design remain open. It is still not clear what is the proper mechanism to gather this type of data. Further experimentation should include incentives and rewards for those users that provide good queries.

The goal of this paper was to explore alternative ways of collecting queries and information needs in a more explicit way. This approach could be useful when constructing data sets to evaluate specific search scenarios. Another utility of the annotation is to help verifying relevance assessments.

6. REFERENCES

[1] Cyril Cleverdon, Jack Mills, and Michael Keen. Factors Determining the Performance of Indexing Systems; Volume 1: Design. Cranfield, 1966.

[2] Cyril Cleverdon, Jack Mills, and Michael Keen. Factors Determining the Performance of Indexing Systems; Volume 1: Design, Part 2. Appendices. Cranfield, 1966.

[3] Edith Law, Anton Mityagin, and Max Chickering. Intentions: A Game for Classifying Search Query Intent. In *Proc. of CHI*, 2009.

[4] Community Query Log Project, UMass, 2010. `http://lemurstudy.cs.umass.edu/`

[5] Fabrizio Silvestri. Mining Query Logs: Turning Search Usage Data into Knowledge. *Foundations and Trends in Information Retrieval* 4(1-2): 1–174, 2010.

[6] Guido Zuccon, Teerapong Leelanupab, Stewart Whiting, Emine Yilmaz, Joemon Jose, and Leif Azzopardi. Crowdsourcing Interactions: Using Crowdsourcing for Evaluating Interactive Information Retrieval Systems. *Inf. Retr.* 16(2): 267–305, 2013.