Using Toolbar Data to Understand Yahoo! Answers Usage*

Giovanni Gardelli
Yahoo! Europe Ltd.
125 Shaftesbury Av.
London, WC2H 8AD, UK
gardelli@yahoo-inc.com

Ingmar Weber
Yahoo! Research Barcelona
Avda. Diagonal 177, 8th floor
08018 Barcelona, Spain
ingmar@yahoo-inc.com

ABSTRACT

We use Yahoo! Toolbar data to gain insights into why people use Q&A sites. We look at questions asked on Yahoo! Answers and analyze both the pre-question behavior of users as well as their general online behavior. Our results indicate that there is a one-dimensional spectrum of users ranging from "social users" to "informational users" and that web search and Q&A sites complement each other, rather than compete. Concerning the pre-question behavior, users who first issue a question-related query are more likely to issue informational questions, rather than conversational ones, and such questions are less likely to attract an answer. Finally, we only find weak evidence for topical congruence between a user's questions and his web queries.

Categories and Subject Descriptors

H.1.2 [**Information Systems**]: User/Machine Systems—*Human factors*

Keywords

community question answering sites, web search, conversational vs. informational, Yahoo! Answers

1. INTRODUCTION

Users have a variety of tools for seeking information online. They can consult web search engines but they can also seek help via social networking sites or ask questions on Q&A sites. We use toolbar data to understand better why people submit questions to Q&A sites. By gathering browsing information for anonymous users we can not only observe the online behavior preceding the creation of a new question online, but we can also construct general user profiles. Understanding question asking behavior on Q&A sites is important as it allows web search engines to better understand alternative search strategies and take a more "social" approach to addressing information needs.

*This research was partially supported by the Torres Quevedo Program of the Spanish Ministry of Science and Innovation, co-funded by the European Social Fund, and by the Spanish Centre for the Development of Industrial Technology under the CENIT program, project CEN-20101037, "Social Media" `http://cenitsocialmedia.es/`.

WWW 2012 Companion, April 16–20, 2012, Lyon, France.
ACM 978-1-4503-1230-1/12/04.

2. RELATED WORK

Harper et al. in [1] describe how to categorize questions in a informational vs. conversational taxonomy through machine learning. This taxonomy is a simpler form of the one proposed in [2] but, due to its simplicity and the categorization accuracy, we apply the binary distinction in our analysis. Relevant bibliography around Q&A behavior is usually focused on social networks, but some results are also applicable to Q&A sites [3]. While it is clear that search and Q&A can be perceived as competitors or complements, it is unclear when these perceptions change and why. Our work sheds some light on this issue by integrating both pre-question search behavior and general online behavior.

3. DATA SET

We used anonymous data collected through the Yahoo! Toolbar from mid-June 2010 to mid-July 2011. We excluded users with less than 1,000 or with more than 1,000,000 page views, or users whose toolbar language was not English. In the end, we used 27,262 distinct users who asked at least one question on Yahoo! Answers (Y!A).

To obtain general user profiles, we classified a subset of URLs into five categories as follows. Q&A page view (Y!A or Wiki Answers), Social page view (Facebook, Myspace or Orkut), Knowledge page view (Wikipedia, *.edu or *.ac.uk), Web search page view (Google, Yahoo! or Bing), and clicked search result page view (referrer was web search).

We also looked at whether a question was preceded (during 10 minutes) by a related web search query. We deemed a (question, web search query) pair related if (i) they were classified as having the same Y!A topic (see below), or if (ii) their Jaccard string token similarity (after normalization and removing stopwords) was $\geq .25$. For preceding web queries, we also looked at whether at least 100 seconds passed after the result page view before another page view as such "long" clicks are better indicators of search success.

For each user, we further categorized up to 1,000 web queries into one of the 26 first level Yahoo! Answers topics. The classifier works by issuing the normalized input string to the Yahoo! Answers Search API and doing a rank-weighted majority voting on the categories returned.

We trained a machine learning algorithm to classify questions into informational vs. conversational. To obtain labeled data, we sampled 500 question instances from our data set; each of these instances was labeled by two judges. There were 265 informational questions, 202 conversational ones, 32 split cases and one ignored case (non-English). A total of 467 labeled questions was then used to train an SVM with

a combination of token uni- and bigrams as features. The trained classifier has 10-fold CV accuracy of 76%, considerably higher than the 57% for a trivial classifier.

4. BASIC ANALYSIS

We wanted to analyze if general web usage is a predictor of question asking behavior. To obtain use profiles we normalized the total page views on Q&A, social, knowledge and web search pages respectively by dividing their count by the total number of page views recorded for that user. Using these fractions, we also bucketed users' questions into 10 percentiles. We then computed regressions to test for correlation between the percentiles of one variable, and the mean of the percentiles of all the other variables. In general there is a *positive* correlation between the usage of Q&A sites and (i) web search percentiles and (ii) knowledge site percentiles (respectively: y=0.213, R^2=0.848; y=0.225, R^2=0.828), indicating Q&A sites are *not* a replacement for search engines or knowledge sites. For social network usage the correlation is negative (y=-0.225, R^2=0.828).

As we observed a strong correlation between the type of question asked and the presence of a related search, we considered four categories: (i) no related searches observed (aware users), (ii) presence of related searches but no clicks on results (discouraged users), (iii) presence of related searches and short clicks on results (failed users), and (iv) long clicks on results (integration users). Fig. 1 shows that if a user had already consulted a web search engine, his question is less likely to attract answers. We also see that questions without a preceding related web query are more conversational, potentially because their information need cannot be satisfied by a search engine.

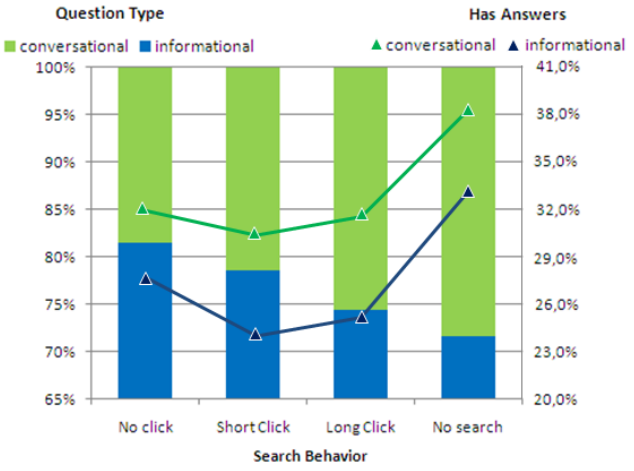

Figure 1: Yahoo! Answers questions are split according to users' pre-question behavior.

Though general pre-question search activity is an indicator for informational questions, questions preceded by related web queries and long clicks have a comparatively high fraction of conversational queries. This might indicate an attempt to integrate information found with users' opinions or suggestions. We also observed a correlation between the presence of a knowledge view before the question posting and the answer probability. On Y!A, given the type of question, there is a much lower chance to receive an answer if there is a knowledge page view right before asking (26.3% vs. 36.8% response rate for conversational questions; 21.8% vs. 31.4% for informational questions).

5. TOPICAL PREFERENCE

Do people ask questions about the same topics they search for? To answer this question we took several approaches. First, we looked at the probability of observing a matching topic pair when one topic is generated according to the user's web search topic distribution and the other topic is (i) also generated according to this distribution, or (ii) is the topic of the user's asked question. Concretely, let p_t^i be the web search topic distribution across topics t for the user pertaining to question instance i. Let $t(i)$ be the topic of this question instance. Then for each given i we compute both (i) $p_{ss}^i = \sum_t p_t^i \cdot p_t^i$ and (ii) $p_{sq}^i = \sum_t p_t^i \cdot 1_{t(i)=t} = p_{t(i)}^i$. As for many instances i we have $p_{ss}^i > p_{sq}^i$ this indicates that users do *not* ask about topics they frequently search for as the match probability is smaller by random chance.

type	#q	$p_{ss}^i > p_{sq}^i$	$p_{sQ}^i > p_{sq}^i$	$p_{Sq}^i > p_{sq}^i$
inf	8,136	79.7%	49.8%	40.7%
conv	2,477	86.5%	48.1%	43.5%

Table 1: Analysis showing the results of the three different tests about users' topical preference.

Second, we corrected for the fact that topics asked online do not follow the same distribution as topics searched for. For example adult topics are prominent in search but banned from Q&A sites. Hence, we looked at the topic match probability when one topic is generated according to p_t^i and the other topic is (i) generated according to the general question topic distribution on Y!A or (ii) is the topic of the user's asked question (as before). Concretely, let p_t^i and p_{sq}^i be defined as before. Define p_t^s to be the question topic distribution across topics t for site s. For a question instance i let $s(i)$ be the site pertaining to that instance. Then for each instance i we compute $p_{sQ}^i = \sum_t p_t^i \cdot p_t^{s(i)}$ in addition to the p_{sq}^i as before. Now we have that $p_{sQ}^i \approx p_{sq}^i$, indicating little influence of personal search history once a general "bias" is taken into account.

Finally, we also corrected for the global topic differences by looking at the probability of observing a matching topic pair when one topic is the topic of the pertaining question and the other topic is (i) sampled from a general web search topic distribution, or (ii) is sampled from the user's web search topic distribution. Case (i) pertains to p_{Sq}^i defined analogously to before and Case (ii) pertains to p_{sq}^i. With this correction, users on Yahoo! Answers tend to ask about their more frequent web search topics.

6. REFERENCES

[1] F. M. Harper, D. Moy, and J. A. Konstan. Facts or friends? Distinguishing informational and conversational questions in social Q&Asites. In *CHI*, pages 759–768, 2009.

[2] M. Harper, J. Weinberg, J. Logie, and J. A. Konstan. Question types in social Q&Asites. *First Monday*, 15(7), 2010.

[3] M. R. Morris, J. Teevan, and K. Panovich. What do people ask their social networks, and why?: a survey study of status message Q&Abehavior. In *CHI*, pages 1739–1748, 2010.