Why do you ask this?*

Using Toolbar Data to Identify Common Patterns of Q&A Users

Giovanni Gardelli
Yahoo! Inc. Barcelona
gardelli@yahoo-inc.com

Ingmar Weber
Yahoo! Research Barcelona
ingmar@yahoo-inc.com

ABSTRACT

We use Yahoo! Toolbar data to gain insights into why people use Q&A sites. For this purpose we look at tens of thousands of questions asked on both Yahoo! Answers and on Wiki Answers. We analyze both the pre-question behavior of users as well as their general online behavior. Using an existing approach (Harper et al.), we classify questions into "informational" vs. "conversational". Finally, for a subset of users on Yahoo! Answers we also integrate age and gender into our analysis.

Our results indicate that there is a one-dimensional spectrum of users ranging from "social users" to "informational users". In terms of demographics, we found that both younger and female users are more "social" on this scale, with older and male users being more "informational".

Concerning the pre-question behavior, users who first issue a question-related query, and especially those who do not click any web results, are more likely to issue informational questions than users who do not search before. Questions asked shortly after the registration of a new user on Yahoo! Answers tend to be social and have a lower probability of being preceded by a web search than other questions.

Finally, we observed evidence both for and against 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

Nowadays, users have a variety of tools for seeking information online. First and foremost, they can consult web search engines [6, 5, 16]. However, they can also seek help via social networking sites [14] or ask questions on Q&A sites [1]. In this work we use toolbar data to understand better why people submit new questions to Q&A sites. This has potential applications to make Q&A sites more engaging, but also to improve web search, as it gives a clearer picture of different information seeking strategies and what their causes are. In particular, our findings reveal what user types consider Q&A sites as either substitutes or complementary to web search, opening many potential integration possibilities between the two tools in order to make web search more social.

We start by stating several hypotheses (H1 through H13), given in Section 2. To address these hypotheses, we looked at new questions posted to Yahoo! Answers[1] (Y!A) and Wiki Answers[2] (Wiki).

Our main findings are:

- There is a "knowledge user dimension", and use of web search engines, Q&A sites and Wikipedia-like sites mutually reinforce each other, rather than compete. (¬H1)
- The knowledge dimension extends into the pre-question behavior, and knowledge users are more likely to perform pre-question searches and ask informational questions. (H2, H4)
- The inverse of the knowledge dimension is a "social user dimension" with users being more active on social networking sites and asking more conversational questions. (H3)
- Questions asked after related web searches are less likely to obtain an answer, but question after failed searches are not more likely to contain a description. (H5, H8)
- Questions immediately after registration are more conversational than expected, as are questions after web queries with clicks with long dwell time. (¬H10, H6)
- The correlations with demographics are as one would expect, with young and female users asking more conversational questions. (H12, H13)
- Searches on search engines and Q&A sites correlate similarly with other variabless. (¬H7)
- Some weak personal topic congruence can be observed for Yahoo! Answers, but not for Wiki Answers. (¬H9)

The rest of the paper is organized as follows. In Section 2 we list the hypotheses that form the starting point of our work. In Section 3 we discuss work related to Q&A sites in general and to Y!A in particular, but also other work

*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://www.cenitsocialmedia.es/.

CQA'2012, a WWW'2012 workshop, April 17, 2012, Lyon, France
ACM 978-1-4503-1230-1/12/04.

[1] http://answers.yahoo.com/

[2] http://wiki.answers.com/

about information finding strategies. In Section 4 we give more details about the two sites used and explain the data extraction process and the final variables obtained. Our hypotheses are addressed in Section 5, with a general focus on "two or three variables at a time".

2. HYPOTHESES

Our research hypotheses are phrased with an existing "informational" vs. "conversational" question taxonomy [8, 9] in mind that distinguishes between one-good-answer-suffices factual questions and what-do-you-think personal questions.

General behavior-related hypotheses

H1: General activity on Q&A sites indicates less other knowledge seeking activity.

H2: Users with a generally high level of search activity are more likely to search before asking a question.

H3: Users who are generally more social have a higher probability of asking conversational rather than informational questions.

Pre- and post-question-related hypotheses

H4: Conversational questions are less likely to be preceded by related web searches.

H5: Questions asked after a failed search attempt are less likely to obtain an answer than questions where the user asks directly.

H6: Questions asked after successful search attempts are more likely to be conversational than those after failed search attempts.

H7: Pre-question web search queries and pre-question queries issued on Q&A sites serve different purposes.

H8: Users are more likely to add more details to questions for which they first searched and failed.

H9: Users tend to ask questions about topics they also generally do web searches on.

H10: Questions for which the user explicitly registered are more likely of an informational nature.

H11: Questions asked after failed search attempts are less likely to be easy and attract trivial answers.

Demographics-related hypotheses

H12: Conversational questions are more prevalent among younger users than among older users.

H13: Male users are more likely to ask informational questions than female users.

To address our hypotheses, we subsequently tried to map the unobservable, high level features such as "a social user", "a related query" or "a failed search attempt" to observable, low level features. Similarly, the textual hypotheses were ultimately mapped to formal hypothesis tests on things such as the equality of two means or the correlation between two variables (see Section 5).

3. RELATED WORK

Although there has been a significant focus on best answer prediction, few researchers directly investigated user behavior. Gyöngyi et al. [7] showed how the category of content is related to the number of answers the question receives on average, while Adamic et al. [1] showed that the asker/replier overlap, too, depends on the topic. This is related to the informational vs. conversational question taxonomy proposed by Harper et al. in [8] where the authors also describe how to automatically tell the two classes apart through machine learning, using both the topic categorization and keywords as features. This taxonomy is a simpler form of the one proposed in [9] but, due to its simplicity and the categorization accuracy, we apply the binary distinction in our analysis. Similar binary taxonomies that distinguish subjective from objective questions have been proposed by Li et al. in [12][13] where the authors use co-training and POS features for question categorization. However, since both these approaches also rely on the answers to derive the features they are not fully applicable to our dataset. The "conversational" type is related to the selection criteria for best answers studied in [11]. There the authors found that the socio-emotional criterion was particularly prominent for opinion and suggestion questions.

To our knowledge there are few papers that investigate the reasons leading users to ask specific questions on Q&A sites. Relevant bibliography around this topic is usually focused on social networks (especially Twitter and Facebook), but some results are also applicable to Q&A sites. For example, the authors of [14] looked at questions contained in status messages in Facebook, and their findings suggest both that a significant proportion of users consider web search and Q&A sites as alternatives for information seeking, and that a fraction of them actually chose Q&A because of a previously failed search. While it is clear that search and Q&A can be perceived as substitutes, competitors or complements, it is less clear when these perceptions change and why. Our work sheds some light on this issue by integrating both pre-question behavior and general online behavior in the analysis. Evans et al. [3] addressed a similar topic, observing the behavior of eight volunteers. In their experiments they observed the synergies between non-social and social information seeking resources that can lead to better results when combined. Their question of how to identify questions that benefit from social input relates to the distinction between social and informational questions applied in our analysis. Thatcher [15] instead, identifying different search strategies, highlighted how some users even prefer to avoid search engines and choose other portals/sites as starting points for searching. Finally it is important to remember that information seeking depends on many personal factors, such as the self-confidence and experience of a user with a particular tool, and certain choices might even not be rationally explainable. As remarked in [2] many users prefer to call customer support and have a social interaction in order to solve a problem, even there are other, faster ways to look for the same information (such as web search). Moreover, a human interaction is proven to be helpful not only to seek a specific piece of information, but also to better define the problem need itself, or to gather feedback about the information found [4].

4. DATA SET

4.1 Toolbar Data

For our analysis we used anonymous data collected through the Yahoo! Toolbar. Although toolbar users possibly differ from other users of Q&A sites, we believe that our observations are robust enough to hold in general, as we focus on trends and correlations between specific variables.

We used a large sample of toolbar data where, for the users in the sample, we used all existing records from mid-

User	feature	aver.	med.	10%	90%
Q&A	# views	38,940	21,549	4,078	84,154
	% Q&A	0.92%	0.23%	0.06%	1.48%
	% social	28.0%	22.9%	0.53%	64.9%
	% knowledge	0.77%	0.21%	0.02%	1.71%
	% web s.	1.82%	1.03%	0.14%	4.35%
Random	# views	14,231	5,975	1,451	34,927
	% Q&A	0.09%	0.01%	0.00%	0.17%
	% social	25.9%	15.9%	0.06%	68.9%
	% knowledge	0.47%	0.08%	0.00%	0.88%
	% web s.	1.59%	0.69%	0.05%	4.04%

Table 1: Basic statistics comparing our set of 39,289 Q&A users and a random sample of 4,961 users.

June 2010 to mid-July 2011. In the following we refer to an individual toolbar record as a page view or just view. Each page view consists of a timestamp, a URL and an anonymous user identifier, as well as meta data such as the toolbar language, whether a page view was a redirect or not and the referrer (if any). For privacy reasons, URLs starting with *https://* are recorded in truncated form without any dynamic parameters.[3]

For our analysis we excluded all users with less than 1,000 records and with more than 1,000,000 records, or users whose toolbar language was not English. We used two user populations for our analysis. First, a set of 39,289 distinct users who asked at least one question on either Yahoo! Answers (see Section 4.2) or on Wiki Answers (see Section 4.3). And, second, a random subset of 4,961 users to obtain reference statistics for comparison purposes(see Tbl. 1).

To obtain general user profiles, we classified a subset of URLs into five categories listed below using regular expressions on the viewed URL.

- Q&A page view: `answers.yahoo.com`, `answers.com`
- Social page view: `facebook.com`, `myspace.com`, `orkut.com`
- Knowledge page view: `wikipedia.org`, `*.edu`, `*.ac.uk`
- Web search page view: `google.*/search.*?q=query`, `yahoo.*/search.*?p=query`, `bing.*/search.*?q=query`
- Clicked search result page view: referrer was a web search page view

Basic statistics about the distributions can be found in Tbl. 1. Note that a page view can be both, say, a clicked search result and a knowledge page view. For each user, we further categorized up to 1,000 web queries, sampling in chronological order, into Yahoo! Answers topics (Section 4.4).

Apart from constructing general user profiles, we also obtained information about the 10 minutes preceding the event of a user posting a question online. Here we recorded the same categorical page view information as before, but we also added information about whether a user's web searches during that period were related to the question he ultimately asked. To estimate this, we used two notions of "relatedness". One that classifies both web search queries and Q&A questions into the Y!A categories. This is explained in detail in the following section and we refer to it as topic-related. The second approach first normalizes both queries and questions by lower-casting them, removing stopwords[4], uniquing tokens and then requires a Jaccard coefficient of $\geq .25$ to label the two objects as string-related. Both approaches gave similar results and we usually treat them as one group. Note that we did not only identify web search queries but, separately, also queries on the two Q&A sites themselves.

Finally, for web queries, we not only counted the number of result clicks (making use of the referrer information) but also looked at whether at least 100 seconds passed after the result page view before another page view. Such "long" clicks have been observed to be better indicators of search success than shorter ones [10].

4.2 Yahoo! Answers

Yahoo! Answers is the most popular Q&A site on the web. It is structured around three main components.

Ask: Users can ask a short question and are given the possibility to add a detailed description or further information (up to to 5,000 characters). They then choose an appropriate category from a list of automatically proposed ones. Duplicate and near-duplicate questions are allowed and exist. Questions that are left unanswered after a period of 4 days, extensible to a total of 8 days by the asker, are removed from the system and are no longer accessible.

Answer: Users are incentivized to answers questions through "points" for each answer they post, and for each positive judgment those answers attract from other users. Users can only submit one answer per question and, unlike in threaded commenting, they cannot answer to previous answers and they cannot collaborate on a single answer in a wiki-manner.

Vote: Users can vote on answers they consider valuable. This promotes better answers and makes them more visible on the page, while rewarding their contributors.

All activities on Yahoo! Answers require the user to be registered with the site and questions and answers can personally identified. Question attempts by users who were not logged in cannot be tracked. We did however record if the user registered on the Yahoo! network during the 10 minutes preceding the question. For a subset of 4,436 distinct Y!A users we also obtained self-provided basic demographic information, comprising age and gender.[5] In the subsequent analysis, we never used a user's identity but only their age and gender, which were analyzed in aggregate. In total, we used toolbar information related to 27,262 instances of questions being asked on Yahoo! Answers. The activity range was from 21,067 users with a single question instance on Yahoo! Answers, to one user with the maximum of 6 instances.

4.3 Wiki Answers

Wiki Answers is the second most popular Q&A site on the web. It is organized around the following components.

Ask: Users can post a question of length up to 200 characters. If the question or a near-duplicate is already present on the site, a user is redirected there. Otherwise users are provided with a list of similar questions and the option to post the new one to the site. This system tries to avoid duplicate content. Categorization of new questions is optional and several topics can be chosen for the same question.

Answer: Users can improve the current answer or create a

[3] Note that a "user" is a toolbar ID and an individual might use several machines with toolbars, or a single machine might be shared by several individuals.

[4] `http://www.ranks.nl/resources/stopwords.html`

[5] We only obtained this information for a subset of those users who registered before April 2011.

new one. This component is designed to avoid duplication of content, and pushes the users to generate a comprehensive and coherent answer instead of many fragmented ones.

Edit: Users can edit and merge the content of both questions and answers, and they are incentivized through a leveling system to contribute to improve the quality of information on the site.

Asking a question on Wiki Answers does *not* require any registration and all questions are posted anonymously. Answering and other forms of non-trivial edits do, however, require registration. Overall, there are less "community aspects" pertaining to Wiki Answers when compared to Yahoo! Answers. For Wiki Answers a not-yet-successful question attempt in progress cannot be distinguished from a regular search on the site and so we only recorded cases were the new question was finally posted online. In total, we used toolbar information related to 18,015 instances of questions being asked on Wiki Answers. The activity range was from 13,075 users with a single question instance on Wiki Answers, to two users with the maximum of 5 instances.

In our analysis (Section 5), we observed the two sites to have very different characteristics and we report results separately for each of the two sites.

4.4 Topic Classifier

To detect if a question and a query are topically similar we devised a simple topic classifier that classifies these objects into one of 26 first level Yahoo! Answers topic categories, as well as a "Unknown" category. The classifier works by (i) first casting the input string (either a web search query or a question) to lower case, (ii) tokenizing it on non-word characters, (iii) removing frequent stopwords[4], (iv) uniquing the tokens and (v) issuing the concatenated, distinct tokens to the Yahoo! Answers Search API. Up to 10 search results were then retrieved and for each one of them the first level topic category was recorded. The result at position i then "voted" for its category with weight $21 - i$. If no search results were returned and there were more than three distinct non-stopword tokens, we removed the last token from the input string and repeated the process. If there were already no more than three distinct tokens but no search results then the input string was classified as "Unknown". This often happened for web queries with either typos or URLs, as well as for web queries in non-English languages. Note that for Y!A questions that were still online, i.e. the user went through the whole asking process and the question attracted at least one answer (as otherwise it will be removed), the question to be classified was returned as the most relevant question and contributed most in the voting process. Tbl. 3 shows the topic distribution according to the classifier for questions asked on Yahoo! Answers and Wiki Answers (in the second column) and for web searches done by the Q&A users (in the third column). All topic distributions are micro-averaged across question instances, i.e. more active users contribute more.

	%-age	%-age	%-age	
Topic	questions	queries	inf.	conv.
Arts&Humanities	3.3 / 7.4	3.3	81.0	19.0
Beauty&Style	2.6 / 1.1	3.9	22.7	77.3
Business&Finance	4.6 / 3.6	3.6	92.5	7.53
Cars&Transport.	4.2 / 8.1	2.3	90.2	9.8
Comput.&Internet	10.1/ 3.2	9.3	96.1	3.9
Consumer Electr.	4.0 / 1.4	2.1	95.1	4.9
Educ.&Reference	4.8 / 8.9	2.3	74.2	25.8
Entert.&Music	6.6 / 7.0	10.2	84.8	15.2
Fam.&Relationsh.	8.0 / 5.1	3.0	6.2	93.8
Games&Recreation	3.0 / 1.8	3.8	94.5	5.5
Health	7.5 / 6.0	2.8	63.8	36.2
Pets	3.1 / 2.4	1.5	23.3	76.7
Polit.&Governm.	4.8 / 8.8	3.5	86.2	13.9
Pregn.&Parent.	2.6 / 2.1	1.2	7.3	92.7
Science&Mathem.	5.6 / 11.5	2.3	98.0	2.0
Society&Culture	5.4 / 8.1	3.8	62.1	37.9
Sports	2.5 / 3.3	3.2	95.8	4.2
Travel	2.0 / 2.7	2.2	80.4	19.6
Yahoo! Products	6.6 / 0.4	1.2	98.2	1.8
Unknown	3.8 / 2.4	29.5	93.2	6.8

Table 3: Topic distribution for (i) Y!A (2nd column, 1st number), (ii) Wiki (2nd column, 2nd number) and (iii) web searches done by the Q&A users in our data set. The last two columns give a breakdown for questions on Y!A into informational and conversational. Only topics accounting for at least 2% on Y!A, Wiki or in web search are shown.

4.5 Informational vs. Conversational

Users can have different motivations for asking and one basic taxonomy, as proposed in [8], is the distinction between *informational* and *conversational* questions. Informational questions aim at a single good answer and tend to be more factual. An example would be "How can I change the font size in a table in Latex?". Conversational questions on the other hand are better satisfied by a number of answers and they tend to evolve more around opinions. An example would be "What do you think of the dominance of Microsoft products?".

To obtain labeled data to train a machine learning algorithm, we sampled 500 question instances from our data set, both for Y!A and for Wiki. Each of these 1,000 instances was presented to two judges who labeled them as either informational or conversational. In less than one percent of the cases, questions were disregarded from consideration as they were not sensible or in a non-English language. For Y!A there were 265 informational questions, 202 conversational ones, 32 split cases and one ignored case. The corresponding numbers for Wiki were 358 informational, 95 conversational, 41 split and 6 ignored. Only cases where the combined decision was not a split were used in the training phase. Note that this distribution is comparable to [8], both in terms of overall ratio of informational questions to conversational questions, and also in terms of the percentage of split cases. One observes that there is a substantial difference between the two sites regarding the relative proportion of question types.

A total of 920 labeled questions was then used to train a classifier. As features we used a combination of token uni- and bigrams. This combined list was then sorted by frequency and we used the 500 most frequent of them. To this feature set we added the question topic, as output by our classifier, as well as the site identifier (Y!A or Wiki). This is the same feature set used previously in [8].

			pre-q web search			pre-q Q&A search				answered only	
site	type	# q	all	string	topic	all	string	topic	answered	descr.	url
Y!A	inf.	20,063	40.6%	22.9%	23.4%	16.1%	10.4%	10.1%	30.9%	66.9%	21.7%
Y!A	conv.	7,199	34.4%	15.7%	17.3%	14.4%	8.4%	8.8%	36.2%	77.6%	12.5%
Wiki	-	18,015	24.9%	13.2%	10.4%	-	-	-	17.6%	-	1.2%

Table 2: Basic statistics about the questions (and the answers) posted on the two sites used for our analysis.

Site	accuracy	P	R	F1
Y!A	75.7 (56.7)	73.1 (56.7)	89.9 (100)	80.4 (72.4)
Wiki	79.8 (79.0)	81.1 (79.0)	97.1 (100)	88.3 (84.0)

Table 4: 10-fold cross validation performance results for the binary informational ("yes") vs. conversational ("no") classification task. The performance of a constant classifier is given in parentheses.

We used SVM^{perf}[6] to train a support vector machine with a linear kernel function for the classification task. Most parameters were kept at their default values but we increased the value of the constant c governing the trade-off between training error and margin ($c = 1$, $w = 9$, $o = 2$, $t = 0$, $p = 1$).

For Yahoo! Answers the trained classifier has 10-fold CV accuracy of 76%, considerably higher than the 57% for a trivial classifier. This performance is still short though of the performance reported in [8], most likely as the topic label was the output of another classifier. However, for our general analysis we focus on trends, relative comparisons and correlations. These would likely be stronger for a better classifier, but are unlikely to change directions.

For Wiki Answers, however, we did not manage to improve over the trivial baseline. Correspondingly, we never applied the informational/navigational distinction to Wiki Answers. Note that Wiki Answers differs from large Q&A sites in its lack of emphasis on the community aspects and, for this reason, it had also been dropped from consideration in previous work [9, 8].

4.6 Complete List of Variables

Here we give a complete list of the variables we extracted and used in our analysis on a per-question basis. Note that not all features were present for all questions and we did not use all the features for all the different analyses we did.

4.6.1 General Behavioral Variables

This group of variables describe a user's general online behavior, independent of a particular question instance.

- Total number of page views
- %-age (Q&A / social / knowledge) page views
- %-age web search query page views
- %-age web search result page views
- %-age web searches classified by topic
- points on Y!A (Y!A only)
- # questions asked according to user's profile (Y!A only)
- # answers given according to user's profile (Y!A only)

4.6.2 Pre-Question Variables

These variables describe the user behavior during the 10 minutes preceding the question instance.

- has newly registered on Yahoo!
- # web search queries
- # (string / topic)-related web search queries
- # queries on Q&A sites
- # (string / topic)-related queries on Q&A sites
- # clicked web results
- # long clicks on web results
- # (Q&A / social / knowledge) page views

4.6.3 Post-Question Variables

These variables pertain to a particular question but are independent of the pre-question user behavior.

- has an answer
- question topic
- question type (inf. vs. conv.)
- question has description (Y!A only, only if answered)
- first listed answer contains URL[7]
- number of answers received (Y!A only)

4.6.4 Demographic Variables

- gender (subset of Y!A only)[8]
- age (subset of Y!A only)

All of our analysis was anonymous and performed in aggregate. Except where stated otherwise, we always applied a per-question approach and *not* on a per-user approach. That is to say if one user has several question instances his user-specific features are replicated. We chose this approach as the vast majority of question instances came from users with a single question instance in the first place. Furthermore, advertising is typically sold on a per-instance basis and not a per-user basis. So we were more interested in "why did this question instance appear?" rather than "why did this user ask a question?".

5. BASIC ANALYSIS

In this section we address our hypotheses (Section 1) by looking at two to three variables at a time and quantifying how they behave.

[6] `http://www.cs.cornell.edu/People/tj/svm_light/svm_perf.html`

[7] For Wiki Answers there is at most one answer. For Y!A if there is a best answer it is the first listed answer.

[8] A small set of these users also asked questions on Wiki Answers, but the size was too small for significant comparisons.

5.1 General Behavioral Variables

We wanted to analyze if general web usage is a predictor of question asking behavior (Hypotheses H1, H2 and H3). To obtain variables indicating relative activity biases 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 (or rather their questions) into 10 percentiles.

As a first analysis, we wanted to observe if there is a correlation between different variables about general web usage (H1). We computed several regressions to test for correlation between the percentiles of one variable, and the mean of the percentiles of all the other variables. Using the percentiles gave stronger correlations between the variables than the raw fractions as their interdependence is nonlinear.

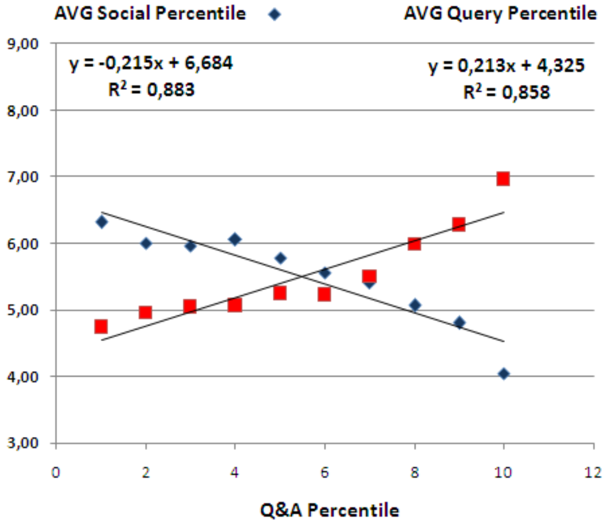

Figure 1: Questions are split into 10 percentiles of fraction of Q&A site page views on the X-axis. In blue are the averages for the corresponding percentiles for social and in red for web search views.

In general there is a positive correlation between the usage of Q&A sites and query-search-knowledge percentiles (respectively: y=0.213, R^2=0.848; y=0.233, R^2=0.916; y=0.225, R^2=0.828), indicating H1 is false. For social network usage the correlation is negative (y=-0.225, R^2=0.828).

Subsequently we tested H3: are users who ask informational questions characterized by a different web usage than users asking conversational questions? Tbl. 5 shows that bigger differences are observed regarding Q&A social views, showing a correlation between the usage of those two categories of sites and the type of question asked on Y!A. Nevertheless, also the usage of knowledge websites and the searching disposition are related to the type of question asked on Y!A. Our findings confirm implications of [15] about how previous knowledge influences the searching behavior.

5.2 Pre-Question Variables

In this section we investigate how the behavior, in particular search behavior during the ten minutes before asking the question, relates to things such as the question type or the probability of being answered. After observing a strong correlation between both the type of question asked (inf. vs. conv.) and the presence of a related search, we decided to split the different pre-question behaviors into four categories: (i) no related[9] 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). Note that the presence of a related query and the presence of long clicks are *separate* variables and clicks could be on unrelated search results.

	Av. percentile of views			
type	Q&A	knowl.	social	web s.
inf.	5.62	5.54	5.23	5.80
conv.	6.21	5.43	5.75	5.54
total	5.77	5.51	5.37	5.73

Table 5: Conversational questions tend to be asked by users in a higher activity percentile for Q&A and social sites. The breakdown is for questions on Y!A and the last row shows the overall average percentile for this set.

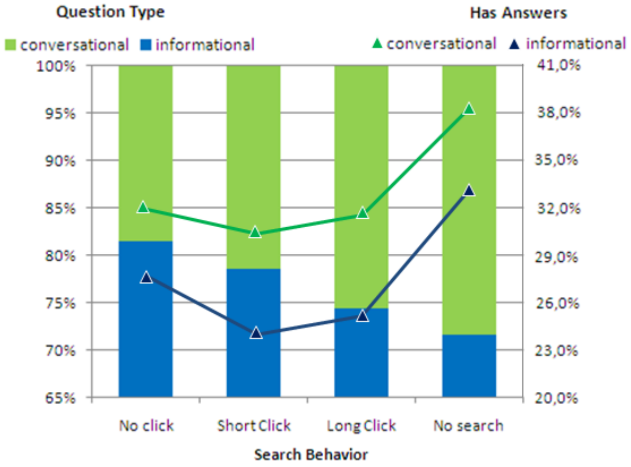

Figure 2: Yahoo! Answers questions are split according to users' pre-question behavior.

We tested first H4, H5 and H6, correlating the type of question asked to the pre-question behavior. As seen in Fig. 2, the different behaviors regarding web search are correlated both with the type of question asked on Y!A and its chance to be answered.

The arguably more difficult cases, where the user had already consulted a web search engine, are less likely to attract answers. The same analysis on Wiki, without the question type, has the same statistically significant trend (answers probability: no clicks 14.6%, short clicks 16.5%, long clicks 16.7%, no search 17.9%). Fig. 2 also shows that questions without a preceding related web query are more conversational, potentially because their askers know that their information need cannot be satisfied by a search engine. Furthermore, there is evidence for H6.

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

[9] Either string- or topic related.

fraction of conversational queries. This might indicate that the user found some information and is now trying to integrate it with users' opinions or suggestions.

The same trends (related search ⇒ more informational and smaller answer probability) also hold for pre-question searches on Q&A sites, refuting H7. However, it trivially fails for Wiki, where the user is obliged to first issue his question as a query.

We tested also if the general behavior is related to the pre-question behavior: given that users with more social activity search less than average, and users with knowledge activity search more than the average, we did not observe any relevant difference for the presence of related queries.

The last analysis related to the behavior in the last 10 minutes concerns a correlation between the presence of a knowledge or social view before the question posting. 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). No significant correlations have been observed with presence or absence of social views in the prior 10 minutes. Also given an informational question, if the user searched on the Q&A site for similar things before, then the probability to find a link in the answer is slightly lower (19.2% vs. 22%), potentially indicating a less trivial answer, supporting H11.

On Wiki, however, both the correlations with knowledge and social views are significant: given a knowledge view right before the question, the answer rate drops from 17.9% to 14.7%. The correlation with social views is in the opposite direction: the answer probability increases from 17.7% to 21.2%, possibly due to a higher fraction of social questions following such views. This relevancy might be caused by the fact that we did not split this analysis by type of question, so the two different behaviors might just reflect the type of question asked, on which the answer rate depends, more than the answer rate itself.

5.3 Post-Question Variables

Here we discuss the interplay of variables that are question-dependent (unlike in Section 5.1 and 5.4) but independent of the pre-question behavior (unlike in Section 5.2).

As on Yahoo! Answers conversational questions are more likely to get an answer than informational ones (see Tbl. 2 and Fig. 2), we computed the following analysis considering only the answered questions.

If an informational question on Y!A is answered, the answer is more likely to contain a URL, with a probability of 21.7% against 12.5% for conversational ones. This is an indication that such questions are more amendable to pointers to factual information, whereas conversational questions are best addressed by the feedback from other users. Also question descriptions on Y!A are more prominent for conversational questions (77.6%) than for informational questions (66.9%) (see Tbl. 2); but surprisingly the pre-question behavior does not influence significantly the presence of a description given the type of question asked, so we may refute H8. This indicates that the motivation to give a description is rather to stimulate discussion than to give details about a concrete problem at hand. The last two tests confirm the truthfulness of the "informational" and "conversational" definitions initially used to categorize the questions database.

Subsequently we tested also, given the type of question, if there are relevant differences related to the pre-question behavior: we observed a significant difference only for conversational questions, for which aware users got a link in 11.7% of given answers, discouraged users got it on 19.8% of cases, failed users 17.7%, and integration users only 2.3%. These findings support H11, assuming that an answer that contains a link is more "trivial".

The same analysis for Wiki was not significant due to the low total number of links in the answers (only 37). This can be explained by the different design: Wiki aims to provide self-contained answers, while on Y!A users are stimulated to provide a link as reference to their sources.

Finally we looked at the popularity for questions on Y!A with at least one answer. For the subsets of informational and conversational questions 41% and 27% respectively only had a single answer. This difference was even more pronounced for the average number of answers due to cases of conversational questions with more than a dozen of answers. This reconfirms the motivation behind distinguishing between these question types.

5.4 Demographic Variables

Fig. 3 shows differences concerning the question type (inf. vs. conv.) and the question count for different demographic segments. Overall, for questions asked by users with demographic information the average age was 38 years and the male/female split was 39.5%/60.5% on a per-question basis. This indicates that this user group is more female-biased and younger than the population of web search users (see [16]), at least when weighted by activity.

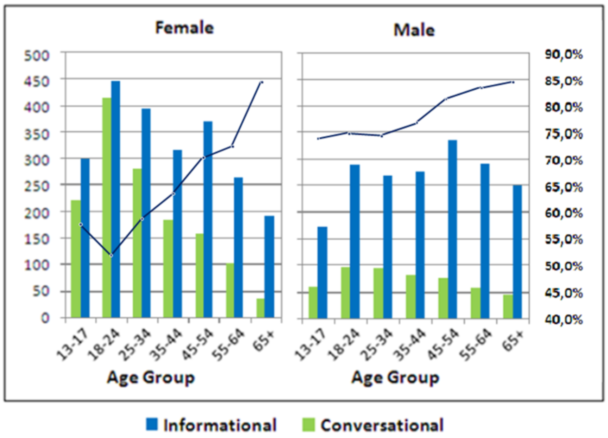

Figure 3: Age gender breakdown for questions on Yahoo! Answers with demographic information, showing counts for informational and conversational questions. Both older users and male users have a higher fraction of informational questions, supporting H12 and H13.

Apart from pre-question searches, we also looked at 135 cases where a user registered for the first time on Y!A during the 10 minutes preceding the question. Such users tend to ask a conversational question more than already registered users. While the fraction of conversational questions from users already registered is 26.3%, it raises to 45.2% for the new users, negating our initial Hypothesis H10. This might indicate that a user's main motivation for registering is the desire to get information he knows he "cannot get elsewhere".

5.5 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^i_t be the 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) $p^i_{ss} = \sum_t p^i_t \cdot p^i_t$ and (ii) $p^i_{sq} = \sum_t p^i_t \cdot 1_{t(i)=t} = p^i_{t(i)}$. If for many instances i we have that $p^i_{ss} > p^i_{sq}$ then this indicates that users usually do *not* ask about topics they frequently search for as the probability of a topic match is smaller than expected by random chance. As can be seen in Tbl. 6, the fraction of such cases is large, giving a first indication against H9.

site	type	#q	$p^i_{ss} > p^i_{sq}$	$p^i_{sQ} > p^i_{sq}$	$p^i_{Sq} > p^i_{sq}$
Y!A	inf	8,136	79.7%	49.8%	40.7%
Y!A	conv	2,477	86.5%	48.1%	43.5%
Wiki	-	4,481	89.3%	55.9%	51.1%

Table 6: Analysis showing the results of the three different tests about users' topical preference. Only users who issued at least one web query were considered.

Second, we tried to find out if the question topics are biased towards a user's search topics once we correct for the fact that, generally, the topics asked online do not follow the same distribution as the topics searched for. This can be because people generally ask about and search for different topics (and adult topics are prominent in search logs but banned from Q&A sites) or it can be because our topic classifier might have a bias and work differently for the (longer) Q&A questions compared to the (shorter) web queries. Hence, 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 p^i_t and the other topic is (i) generated according to the general question topic distribution for the respective site or (ii) is the topic of the user's asked question (as before). Concretely, let p^i_t and p^i_{sq} be defined as before. Define p^s_t to be the question topic distribution across topics t for site s (either Y!A or Wiki).[10] For a question instance i let $s(i)$ be the site pertaining to that instance. Then for each instance i we compute $p^i_{sQ} = \sum_t p^i_t \cdot p^{s(i)}_t$ in addition to the p^i_{sq} as before. Now if for many i we have that $p^i_{sQ} > p^i_{sq}$ then this indicates that users do *not* ask about topics they also search for, even when a general "bias" is taken into account. Here Tbl. 6 gives weak evidence that H9 holds for Y!A but not for Wiki.

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^i_{Sq} defined analogously to before and Case (ii) pertains to p^i_{sq}. With this correction, users on Yahoo! Answers tend to ask about their more frequent web search topics supporting H9 and in line with [7]. Interestingly, there is still evidence for a de-personalized behavior on Wiki Answers, refuting H9 for this site. We hypothesize that this can be explained by Wiki's policy to strongly discourage (near-)duplicate questions hence forcing users into niche topics.

[10] For Tbl. 6 we further conditioned on the question type.

6. REFERENCES

[1] L. A. Adamic, J. Zhang, E. Bakshy, and M. S. Ackerman. Knowledge sharing and Yahoo! Answers: everyone knows something. In *WWW*, pages 665–674, 2008.

[2] K. Ehrlich and D. Cash. Turning information into knowledge: Information finding as a collaborative activity. In *DL*, pages 119–125, 1994.

[3] B. Evans, S. Kairam, and P. Pirolli. Do your friends make you smarter?: An analysis of social strategies in online information seeking. *IPM*, 46(6):679–692, 2009.

[4] B. M. Evans and E. H. Chi. Towards a model of understanding social search. In *CSCW*, 2008.

[5] N. Ford, D. Miller, and N. Moss. Web search strategies and human individual differences: A combined analysis. *JASIST*, 56(7):757–764, 2005.

[6] N. Ford, D. Miller, and N. Moss. Web search strategies and human individual differences: cognitive and demographic factors, internet attitudes, and approaches. *JASIST*, 56(7):741–756, 2005.

[7] Z. Gyongyi, G. Koutrika, J. Pedersen, and H. Garcia-Molina. Questioning Yahoo! Answers. In *QAWeb2008@WWW*, 2008.

[8] 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.

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

[10] A. Hassan, R. Jones, and K. Klinkner. Beyond dcg: User behavior as a predictor of a successful search. In *WSDM*, pages 221–230, 2010.

[11] S. Kim, J. S. Oh, and S. Oh. Best-answer selection criteria in a social q&a site from the user-oriented relevance perspective. *JASIST*, 44(1):1–15, 2007.

[12] B. Li, Y. Liu, and E. Agichtein. Cocqa: co-training over questions and answers with an application to predicting question subjectivity orientation. In *EMNLP*, pages 937–946, 2008.

[13] B. Li, Y. Liu, A. Ram, E. V. Garcia, and E. Agichtein. Exploring question subjectivity prediction in community qa. In *SIGIR*, pages 735–736, 2008.

[14] 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.

[15] A. Thatcher. Information-seeking behaviours and cognitive search strategies in different search tasks on the WWW. *ERGON*, 36(12):1055–1068, 2006.

[16] I. Weber and A. Jaimes. Who uses web search for what? And how? In *WSDM*, pages 15–24, 2011.