2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining

# Finding the Right Social Media Site for Questions

Zhen Yang*, Isaac Jones†, Xia Hu‡, Huan Liu†
*College of Computer Science, Beijing University of Technology, Beijing, China 100124
†Computer Science and Engineering, Arizona State University, Arizona, USA 85281
‡Computer Science and Engineering, Texas A&M University, College Station, TX, USA 77843-3112
Email: *yangzhen@bjut.edu.cn, †{ipjones, huan.liu}@asu.edu, ‡hu@cse.tamu.edu

***Abstract*—Social media has become a part of our daily life and we use it for many reasons. One of its uses is to get our questions answered. Given a multitude of social media sites, however, one immediate challenge is to pick the most relevant site for a question. This is a challenging problem because (1) questions are usually short, and (2) social media sites evolve. In this work, we propose to utilize topic specialization to find the most relevant social media site for a given question. In particular, semantic knowledge is considered for topic specialization as it can not only make a question more specific, but also dynamically represent the content of social sites, which relates a given question to a social media site. Thus, we propose to rank social media sites based on combined search engine query results. Our algorithm yields compelling results for providing a meaningful and consistent site recommendation. This work helps further understand the innate characteristics of major social media platforms for the design of social Q&A systems.**

## I. Introduction

Social media has become a part of our daily life. Most use it, some are developing a dependence on it, and a few make their living on it. Social networks open up new possibilities for discovering information, sharing ideas, and interacting with others. We post on Blogger, tweet on Twitter, like on Facebook, connect on LinkedIn, watch on YouTube, search on Wikipedia, read on Reddit, ask on Ask, message on WhatsApp, and live on Second Life [1]. It's no surprise that we embrace the power of social networks and try to find answers by asking questions of our virtual friends, and do so every day in our social networks.

Unfortunately, our questions don't always receive satisfactory responses, which occurs partly because the responses are highly variable among social media sites, i.e., the quality of responses is quite different when asked on different social media sites. For example, if we want to ask "How do I create an MS Word document within an Android app?", we should post on Stackoverflow (a Q&A site for professional programmers) rather than Facebook or Twitter. Conversly, we should ask "What's a good Mother's Day gift?" on Facebook or Twitter if you don't think "Teach Yourself Visual C++.Net in 21 Days" would make a good gift. Unlike seeking answers from search engines like Google, researchers [2] found searching in social media is strongly tied to people's natural interactions and not analogous to information seeking in more traditional IR environments. Paul [3] conducted a study of Q&A behavior on Twitter and found that the most frequent questions were rhetorical or factual ones. Morris [4] surveyed the questions people ask to their social networks and found that rather than exclusively using social networking services for entertainment, participants reported using them to find practical information and that the most frequent question types are recommendation and opinion. In short, the highly variable Q&A behavior is partly because: (1) Users, though they often turn to their social networks to fulfill their information needs, only have a rough idea of what exactly their questions are. If we cannot describe our questions clearly, it's difficult to find a good answer. (2) Social media sites are a good place to get some questions answered, as people with similar interests from social networks. However, there are many social media sites of disparate types and the interests of social media sites change with users' interests.

In this sense, if responses are highly variable among social media sites, it's intuitive that we should choose the right social media site before asking a specific question. To remedy this, in this paper, through topic specialization, we can find the most suitable social media site for a question with a clear definition. To do topic specialization, previous research on social media has focused on ***the nature of users***, such as the number of followers, days on Twitter, number of tweets, the frequency of use of Twitter, or ***the nature of social ties***, such as relationship reciprocity, tie strength, dyadic interaction over time, link analysis [2][3][4]. However, though online social Q&A behavior has been discussed thoroughly, not much thought has been given to the exploration of ***the nature of the sites*** themselves. Even though some researchers [5] suggested exploring clusters of social media sites, and their sentiment, there is still much to investigate about the nature of the social sites, which is implied by the social interactions that take place on them. One way to overcome this limitation is to look into the content of social sites, i.e., the messages exchanged among users within a social site's domain.

However, topic specialization through the nature of the social sites is an extremely challenging problem: (1) In most social media Q&A, users' questions are always short since they are not clear of what exactly their questions mean. Defining a user's question and expanding it with minimal risk become a urgent problem. (2) With social media on this incredible rise, new social media sites are constantly being created and existing social media sites are constantly changing to match new technology trends which opens up a great challenge, capturing the dynamic of an extremely large number or quickly evolving social media sites in time. (3) For a social media site, we can obtain its content through several sources and methods. However, this content maybe highly conflicted, e.g., the frequency estimation of a single words maybe the same, similar, or totally different. How to combine this highly conflicting content poses a serious challenge.

Meanwhile, we noticed that there are some opportunities. Through the lens of strong semantic knowledge - Wikipedia, and weak sematic knowledge - search engines, we can dynamically capture and understand the content of social sites and rank social sites for a given question by matching the content similarity between questions and social sites, rather than the characteristics of users and their ties. Thus in this way, we provide a framework for topic specialization for a

ASONAM '15, August 25-28, 2015, Paris, France

DOI: http://dx.doi.org/10.1145/2808797.2809391

given question. Our contributions are summarized as:

- We provide a framework (§2.1) for topic specialization by ranking sites for a given question by matching the content of the question and the site, rather than the characteristics of users and their ties.
- We propose a novel method to understand a user's short question (§2.2). Based on Wikipedia, after extracting keywords of a given question, we can expand each keyword.
- We propose a novel method to explore the nature of social sites (§2.3). Based on the discovered content preference, we can explain the highly variable Q&A behavior (§2.5) among social media sites.
- We propose several estimators to optimally combine search engine query results. Our algorithm yields compelling results when providing a coherent site recommendation when compared to ground truth generated by Wikipedia (§3).

## II. Topic Specialization Through Semantic Knowledge Exploration

There are two requirements we want our topic specialization framework to meet: (1) The model should make a question more specific and (2) The model should dynamically capture the contents of social sites. In order to meet these two criteria, through the lens of strong semantic knowledge - Wikipedia, we can utilize topic specialization as it can make a question more specific. Through the lens of weak semantic knowledge - search engines, we can dynamically represent the contents of social sites. This framework can dynamically capture and understand the content of users' question and social sites, as a result, can provide further understanding of the inherent characteristics of major social media platforms and social Q&A systems.

### A. Problem Statement

The general problem we address can be formulated as a ranking task:
**Input**: Given a specific question, *Q*, and a set *S*, where each $s_i \in S$ is a social media site.
**Output**: A ranked version $S'$ of the social media site set *S* where each $s_i \in S'$ is ranked according to its likelihood to give a response to question *Q*.

We argue that a brief question is only abstract of user's information needs and it is difficult to infer users' actual search intent and interest. Thus we expand each question using Wikipedia with minimal risk because it is an authority. For an extremely large number or quickly evolving social media sites, we can dynamically capture and understand their content through the lens of search engines. In this way, as Figure 1 shows , we can rank social sites for a given question by matching the content between questions and social sites. In general, we rank sites with three phases:

- **Question index**: given a question $q$, we extract its keywords and query them on Wikipedia, then index these keywords as the word frequency vector $W(q) = \{p(w_1|Wiki) \cdots, p(w_m|Wiki)\}$.
- **Site index**: given a social media site $s_i \in S$, $T(s_i, g_j, n) = \{p(w_1|s_i, g_i, n), \cdots, p(w_m|s_i, g_j, n)\}$ is the $k$-dimensional word frequency vector of the top-$n$ indexed pages return by search engine $g_j$ within site $s_i's$ domain.
- **Sites Ranking**: Thus for question $q$, the site set $S$ can be ranked by $D(s_i, q) = \frac{<T(s_i,g_j,n),W(q))>}{\|T(s_i,g_j,n)\| \times \|W(q)\|}$.

The details of the three phases will be discussed in follows.

### B. Question Modeling

Users usually have a rough idea of what exactly their questions are and send a short question to sites. To remedy this problem, as shown in Figure 1a, we expand the users' question with Wikipedia:

- **Step I**: Extract keywords. For each question, extract its top ranked words as keywords. In this work, we select all nouns as keywords.
- **Step II**: Expand keywords. For each keyword, we expand using its Wikipedia article obtained via API interface.
- **Step III**: Vector indexing. Index all returned Wikipedia articles as a question profile vector.

A long Wikipedia-based profile rather than an short question is submitted to Q&A system, which can represent the user's intent more effectively. As a result, a better modeling of users' question can be achieved.

### C. Modeling a Social Media Site

Although the tools to crawl the content of social media sites are available and have been used extensively, to date there is no an effective way to dynamically capture a representative sample of content. To tackle this challenge, we capture and understand the content of social sites through the lens of search engines, which crawl the most popular, or representative content of social media sites. In this way, as shown in Figure 1b, we crawl the social media sites profile:

- **Step I**: Crawl content. For a candidate social site, we obtained its $n$ most popular pages by searching with the empty string and restricting the domain to the subject site (for example, site:stackoverflow.com).
- **Step II**: Vector indexing. Index all returned web pages as the social site profile vector.

Being able to describe the nature of a social site in terms of the most representative content would overcome the limitations of the current representations of social sites. As a result, a better modeling of site can be achieved.

### D. Ranking Sites by Combined Searching

It is intuitive to attempt to choose a suitable site on which to pose a specific question based on the matching between the content of the social site and the question. As shown in §2.1, the cosine distance is used to measure the similarity. However, the similarity measure $D(s_i, q)$ implies that the similarity estimation is varied for different search engines, $g_j$, and pages of the index $n$. Our experiments suggest that every reasonable $n$ value ($n \geq 5$) can work, and results are, insensitive to this choice. Thus $T(s_i, g_j, n)$ can be simplified as $T^j = \{p(w_1^j), \cdots, p(w_k^j)\}$ , with $j$ varied by using different search engines $g_j$.

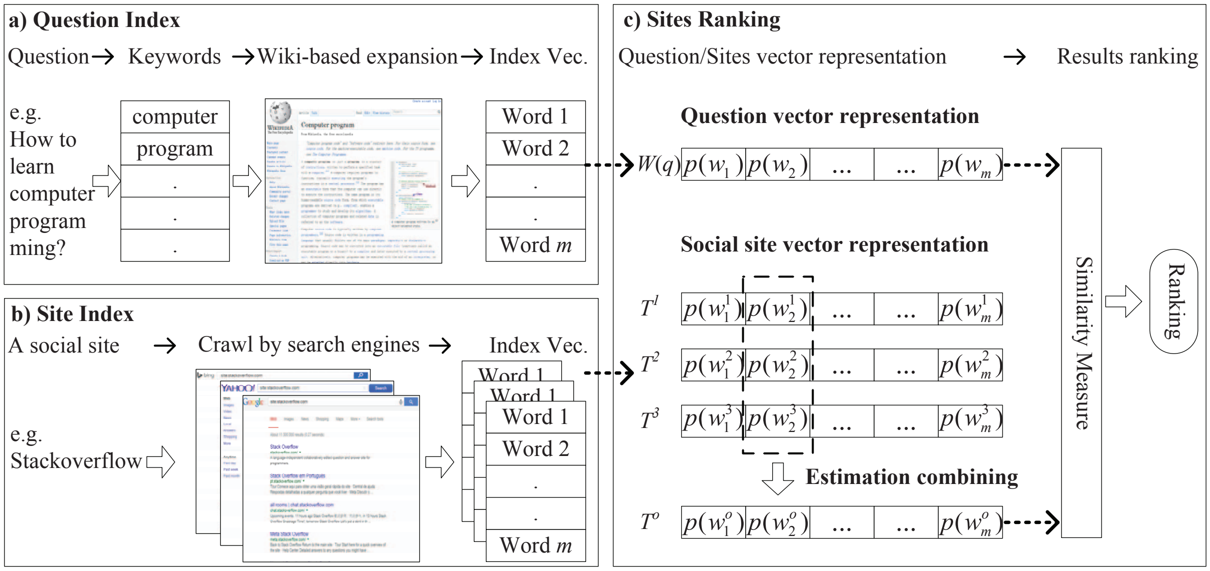


Fig. 1. Diagram of ranking sites for questions. (a) Modeling a users' question, (b) modeling a social media site, and (c) ranking sites by combined searching.

Suppose the optimal combining estimator $T^o = \{p(w_1^o), \cdots, p(w_k^o)\}$, where $p(w_j^o)$ is a combination of the independent sets of probability estimation $\{(p(w_j^1), p(w_j^2), ..., p(w_j^M)\}$ observed by $N$ different search engines. The problem we now face is combing these independent probability estimations. However, since the $p(w_j^i)$ is observed by different search engines, different search engines express their beliefs over the frame according to their techniques, characteristics, update policies, content preferences, etc; we may face conflicting evidence, i.e., for a word, its frequency observation from different search engines may be different.

To overcome this problem, as shown in Figure 1c, we proposed a united framework to combine the evidence obtained by different search engines, and the optimal estimation $p(w_j^o)$ can be denoted as:

$$p(w_j^o) = S(p(w_j^1), p(w_j^2), ...) + C(p(w_j^1), p(w_j^2), ...) \quad (1)$$

where $S(.)$ is the sharing measure function to estimate the common shared belief between multiple sources and $C(.)$ is the conflict function to measure and allocate the conflicting (non-shared) belief. With different sharing measures and conflict allocation strategies, we proposed several evidence combination rules:

(a) Max evidence combination

$$p(w_j^o) = \max_i(p(w_j^i)) \quad (2)$$

(b) Min evidence combination

$$p(w_j^o) = \min_i(p(w_j^i)) \quad (3)$$

(c) Mean evidence combination

$$p(w_j^o) = \frac{1}{M}\sum_i(p(w_j^i)) \quad (4)$$

(d) Dempster-Shafer evidence combination (DS) [6]

$$\begin{aligned} p(w_j^o) &= p(w_j^1) \oplus p(w_j^2) \oplus \cdots \oplus p(w_j^N) \\ &= \frac{1}{1-K} \sum_{\cap_i w_j^i = w_j} \prod_i p(w_j^i) \end{aligned} \quad (5)$$

where $K = \sum_{\cap_i w_j^i = \emptyset} \prod_k p(w_j^i)$.

(e) Yager evidence combination (Yager) [7]

$$p(w_j^o) = \sum_{\cap_i w_j^i = w_j} \prod_i p(w_j^i) \quad (6)$$

(f) Conflict combination (CA):

$$\begin{aligned} p(w_j^o) = & \sum_{\cap_i w_j^i = w_j} \prod_i p(w_j^i) \\ & + q(w_j)(1 - \sum_j \sum_{\cap_i w_j^i = w_j} \prod_k p(w_j^i)) \end{aligned} \quad (7)$$

where $q(w_j) = \frac{\sum_i p(w_j^i)}{\sum_i \sum_j p(w_j^i)} = \frac{\sum_i p(w_j^i)}{M}$.

Combination rules 1 through 6 only measure the shared belief between multiple sources, i.e., are kinds of sharing function $S(.)$, and simply ignores all the conflicting belief or leverages it through a normalization factor, i.e., the conflicting allocation function $C(.) = 0$. This simplification produces wrong results in case of high conflict, e.g., Zadeh's paradox [8]. In rule 7, we try to allocate the conflict probability proportionally. We let the sharing measure function $S = \sum_{\cap_i w_j^i = w_j} \prod_i p(w_j^i)$, in which all probability from different sources are combined in a cumulative manner. Thus the conflicting probability should be $1 - \sum_j \sum_{\cap_i w_j^i = w_j} \prod_k p(w_j^i)$, after the common belief has been taken away. Finally, we allocate the conflicting probability according to the $q(w_j) = \frac{\sum_i p(w_j^i)}{\sum_i \sum_j p(w_j^i)}$, i.e., it is proportional to the ratio of the sum of the probability $p(w_j^i)$ over $i$ to the sum of the probability of whole event.

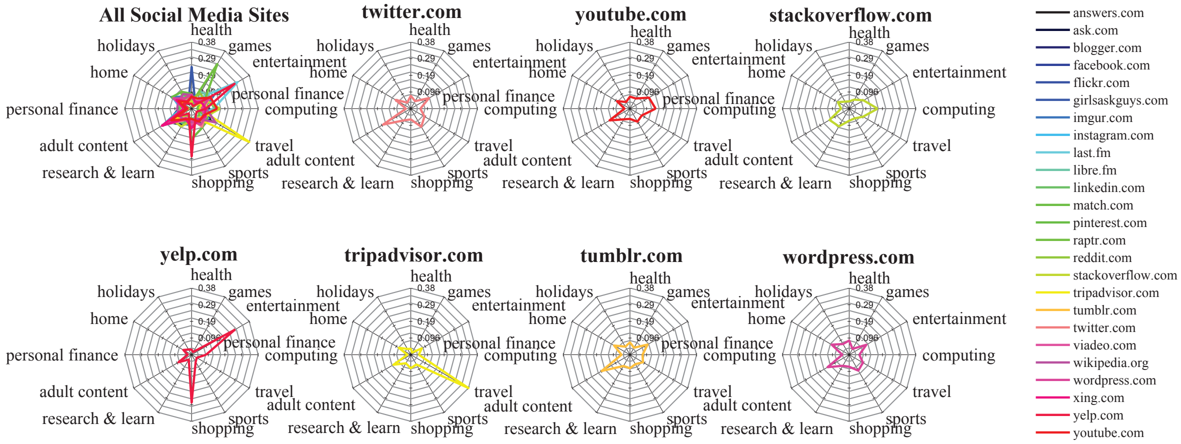


Fig. 2. Distribution of the content of social media sites. Radar diagrams show the similarities between social media site content and topics.

### *E. Content Preference of Social Media*

The novel aspect of the method is the consideration of the nature of social sites, that is, using messages exchanged among users within a social site's domain to judge its appropriateness to answer a specific question, rather than the nature of users and their ties. To better understand the idea behind our algorithm in this section, we explore the content preference of social sites.

First, we select the 12 most asked questions or topics on the Internet [9]. Second, we select 25 typical social media sites, as suggested by many researchers [10]. We roughly divided the social media sites into ten different categories of social media: Blogs, Microblogs, Social Networks, Professional Networks, Media Content Sharing, Collaborative Knowledge Base, Collaborative Filtering, Collaborative Question & Answer, Instant Messaging, and Virtual Social and Game Worlds. As shown in the 3rd column of Table 1, we select some typical sites from each category from the top 200 sites listed on Alexa.com and well known professional sites to explore the nature of their content. For privacy reasons, we cannot probe Instant Messaging Sites, Virtual Social and Game Worlds.

We might investigate whether the social media sites are more inclined to a subset of the topics, i.e., have content preference. With the similar processing phases defined in §2, we can compute the similarity between these topics and social media sites. As shown in Figure 2, radar diagrams are generated showing the similarities between the content of social media sites and the topics. Each diagram is divided into twelve areas of measurement (i.e., the 12 most frequent topics). In each area, the value denotes the similarity between a social media site and a specific topic, where a high value indicates a high degree of similarity. The first radar diagram shows the similarities of all sites (shown in Table 1), and the others are seven selected social media sites, including Twitter, Youtube, Stackoverflow, Yelp, Tripadvisor, Tumblr, and Wordpress.

From Figure 2, we can observe the content preference among social sites. For a social network, such as Yelp and Tripadvisor, strong preferences are shown for their specific business. For more general social sites, such as Twitter, Youtube, and Tumblr, they show bias toward entertainment and adult content, which coincide with our intuition and user experience. Using Twitter as a example, its preference ranking is Adult Content, Entertainment, Sports, Travel, Home, Research & Learning, Shopping, Health, Computing, Games, Holidays, and Personal Finance, which coincides with the survey by S. Paul [3] showing that the most popular topic on Twitter was entertainment (32%). This provides a strong support for utilizing topic specialization before asking, because we shouldn't hope to find a perfect answer for our question in a social media site, which has no content related to our question at all.

## III. Experimental Results

We conduct experiments to evaluate the effectiveness of the proposed framework. Through these experiments, we aim to answer the following questions: 1) How effective is the proposed framework? 2) How does parameter setting affect the performance of the proposed framework?

### *A. Data*

Two data sets were used for experimental evaluation in this work: (1) **Selected Questions**: In Table 1, we selected 10 questions for the top 10 most asked topics on the Internet summarized by [9]. It should be noted that [4] reported that users were hesitant to ask questions about adult content and health, so we ignore these types of questions. (2) **Factoid Q&A Corpus**: In addition, we used the factoid Q&A Corpus [11], which contain 1,714 manually-generated factoid questions and their coreponding answers collected by Carnegie Mellon University and the University of Pittsburgh between 2008 and 2010.

### *B. Experimental Setup*

**Candidate Sites**: As Table I shows, we select 17 well known social media sites from the top 200 sites listed on Alexa.com as the candidate site set $S$. Also, we manually add 8 well known professional social media sites to the candidate site set $S$ for some specific domains, such as Linkedin for job hunting, Match.com for dating, etc. In total there are 25 sites as candidate sites.

**Ground Truth**: (1) For these selected questions, we use the same steps in §2 to generate ground truth. For a specific question, we crawl the Wikipedia articles for its profile per keyword. For these candidate sites, we also crawl their Wikipedia articles as profiles, then rank the sites by the cosine similarity. Since Wikipedia is an authority, these ranking results $S^*$ were

TABLE I. EXPERIMENTAL DATA AND TYPICAL SOCIAL MEDIA SITES.

| Data | Typical Questions | Candidate Social Media Sites |
|---|---|---|
| **Selected Questions** | **Top 10 topics about which people most often asked:**<br>Q1 (Shopping13%): What is a good gift for a girl?<br>Q2 (Entertainment 13%): What music should I listen to?<br>Q3 (Computing 9%): How to learn computer programming?<br>Q4 (Research & learn 9%): How to do my home work?<br>Q5 (Travel 5%): What is the best place to travel<br>Q6 (Games 5%): What is the best electronic game?<br>Q7 (Home 5%): Dos and Don'ts of online dating?<br>Q8 (Sports 3%): When is FIFA world cup 2014?<br>Q9 (Personal Finance 3%): How to do job hunting?<br>Q10 (Holidays 1%): Idea for this valentine's day? | **Top social media sites and their Alexa ranks:**<br>• Blogs: Wordpress(26), Blogger(53)<br>• Microblogs: Twitter(7)<br>• Social networks: Facebook(2)<br>• Professional networks: Linkedin(12), Stackoverflow(45), Tripadvisor(197)<br>• Media content sharing: Youtube(3), Pinterest(25), Instagram(30), Tumblr(39), Imgur(49), Flickr(103)<br>• Collaborative knowledge base: Wikipedia(6)<br>• Collaborative filter: Reddit(50), Yelp(125)<br>• Collaborative Q&A: Answers(200) |
| **Factoid Q&A Corpus** | **About 1,714 manually-generated factoid questions, for example:**<br>Was Abraham Lincoln the sixteenth President of the United States?<br>What is the dominant religion in Ghana?<br>Is Liechtenstein the smallest German-speaking country in the world?<br>Are polar bears excellent swimmers?<br>Why did Grant say "Damn, I had nothing to do with this batte."?<br>What resembles that of the similarly-sized cougar in the Americas?<br>What method is used by Kangaroos to travel?<br>Is a kangaroo on the Australian coat of arms?<br>Did John Adams represent the Continental Congress in Europe?<br>What is the national language of Singapore? | **Well known professional social media sites:**<br>• Music: Last.fm, Libre.fm<br>• Programming: Stackoverflow<br>• Trip: Tripadvisor<br>• Game: Raptr<br>• Dating: Match, Girlsaskguys<br>• Career: Linkedin, Viadeo, Xing |

treated as the ground truth. (2) For the factoid Q&A corpus, their question and answer pairs are already provided, and the articles which answers are extracted from are also provided. For a specific question, we use the article which its answer was extracted from as its profile. For candidate sites, we send them to search engine (e.g. Google) and crawl the top-5 returned pages in the site's domain as profiles, then rank the sites by the cosine similarity. Since the answer articles are manually selected for the specific question, these ranking results $S^*$ were treated as the ground truth.

**Evaluation Metric**: For the ranked version $S'$ of the candidate set $S$, we evaluate the performance by comparing $S'$ and $S^*$ using the top-*n* intersection rate, namely the fraction of the common elements in top $n$ ranking results: $\frac{S'(n) \bigcap S^*(n)}{n}$, where the $S'(n)$ and $S^*(n)$ denote the top $n$ ranking results of $S'$ and $S^*$. Supposing that the top 10 ranked sites in $S^*$ are correct answers of each question, we also can evaluate the top-*n* accuracy rate and average precision (*AP*) [12], which are widely used in IR and keyword evaluation.

### C. Results and Analysis

We compare 9 rule combining settings: Crawling using Google, Yahoo, and Bing, and ranking results using combining rules (a)-(f) defined in §2.4. Table 2 shows the performance of each combining rule on two data sets (10 selected question and and Factoid Q&A corpus) with return pages of search engines as 5, 10 respectively. For each combining rule, we compare the average top-$n$ accuracy rate where the a high value indicates a high precision, the average top-$n$ intersection rate where a high value indicates a high consistency between experimental results and ground truth, and the average precision (AP) instead, where a high value indicates a high average precision on all questions.

To answer the first question of how the proposed framework compares with ground truth, we investigate the effect of combining rules from Table 2:

- **All proposed methods** produce high top-*n* accuracy rate for very low *n* and provide high accuracy rate for top 1 recommendation site in both Q&A corpora, which is valuable in many social applications where users want only one answer. Also, for these selected questions the well accepted sites are ranked in the top of the results as expected. Even better, these sites are listed as the top 1 or 2 sites by all proposed methods.
- **Combining rules** provided a better chance to obtain the higher top-*n* accuracy rate for low $n$ $(n < 5)$ than singleton methods, and the conflict ensemble shows more stable performance on different Q&A corpora. The reason is partly because they have the higher top-*n* intersection rate, i.e., they are more similar to the ground truth. We also see that the conflict allocation, min, and mean ensemble methods have comparable or higher average precision (*AP*) than other methods and produce a higher top-*n* accuracy rate.

To answer the second question of how parameter setting affects the performance of proposed framework, we investigate the effect of return pages and search engines from Table 2:

- **Search engines** produce similar performance, and each search engine has its best result in certain applications. However, combining search engines achieves better and more robust results.
- **Number of returned pages** is not a sensitive parameter. Though results are varied with *n*, however, in a range of very low *n* in practice, our experiments suggest that every reasonable $n$ $(n \geq 5)$ value can work, and our results seem, in general, insensitive to this choice.

## IV. CONCLUSION AND FUTURE WORK

At present, users are using social media in an ad-hoc way without a firm understanding of what the information sources are, what they are useful for, and what the issues involved in using them are. We discussed how to choose a social media site to answer a specific question and why that site is best. We explored this from the angle of the nature of the sites in question, selecting a suitable site for a specific question based

TABLE II. EXPERIMENTAL PERFORMANCE OF PROPOSED MODELS. THE AVERAGE TOP-$N$ ACCURACY RATE, AVERAGE TOP-$N$ INTERSECTION RATE, AND AVERAGE PRECISION (*AP*) OF EACH COMBINING RULE ON TWO DATA SETS (10 SELECTED QUESTIONS AND FACTOID Q&A CORPUS) WITH RETURN PAGES OF SEARCH ENGINES AT 5, 10 RESPECTIVELY.

| Corpus | Returned pages | Method | | The average top-$N$ accuracy rate | | | | | The average top-$N$ intersection rate | | | | | $AP$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | $N$=1 | $N$=2 | $N$=3 | $N$=4 | $N$=5 | $N$=1 | $N$=2 | $N$=3 | $N$=4 | $N$=5 | |
| Selected questions | 5 | Baseline | Google | 0.9000 | 0.7000 | 0.6667 | 0.6500 | 0.6200 | 0.6000 | 0.4000 | 0.4333 | 0.4250 | **0.4200** | 0.6355 |
| | | | Yahoo! | **1.0000** | **0.9000** | 0.7333 | **0.7250** | 0.6400 | 0.7000 | 0.7000 | 0.5000 | 0.4250 | 0.3800 | 0.6784 |
| | | | Bing | 0.8000 | 0.8000 | 0.6000 | 0.6000 | 0.6000 | 0.6000 | 0.6000 | 0.4333 | 0.4000 | 0.4000 | 0.6512 |
| | | Comb. Rule | Max | 0.8000 | 0.7000 | 0.7000 | 0.6000 | 0.6000 | 0.5000 | 0.5500 | 0.5000 | 0.4000 | 0.3600 | 0.6442 |
| | | | Min | 0.8000 | 0.8500 | **0.7667** | 0.7000 | **0.6600** | **0.8000** | 0.6000 | 0.4667 | 0.4250 | **0.4200** | **0.6831** |
| | | | Mean | **1.0000** | **0.9000** | **0.7667** | 0.6500 | 0.6400 | 0.7000 | **0.7500** | **0.5333** | **0.4500** | 0.3800 | 0.6812 |
| | | | D-S | 0.9000 | 0.7000 | 0.7333 | 0.6500 | 0.6200 | 0.7000 | 0.4000 | 0.4333 | 0.4250 | 0.3800 | 0.6411 |
| | | | Yager | 0.9000 | 0.7000 | 0.7333 | 0.6500 | 0.6200 | 0.7000 | 0.4000 | 0.4333 | 0.4250 | 0.3800 | 0.6411 |
| | | | CA | **1.0000** | **0.9000** | **0.7667** | 0.6500 | 0.6400 | 0.7000 | **0.7500** | **0.5333** | **0.4500** | 0.3800 | 0.6801 |
| | 10 | Baseline | Google | 0.8000 | 0.7000 | 0.6667 | 0.6500 | 0.6400 | 0.6000 | 0.4500 | 0.4000 | 0.3500 | 0.4400 | 0.6358 |
| | | | Yahoo! | **1.0000** | **0.9000** | 0.7333 | 0.7000 | 0.6400 | 0.7000 | 0.6500 | 0.4667 | 0.4250 | 0.3800 | 0.6645 |
| | | | Bing | 0.9000 | 0.8500 | 0.6667 | 0.5500 | 0.5800 | **0.8000** | 0.6000 | 0.4667 | 0.4250 | 0.4000 | 0.6799 |
| | | Comb. Rule | Max | 0.8000 | 0.8000 | 0.7000 | 0.6750 | 0.6400 | 0.5000 | 0.6000 | 0.5000 | 0.4500 | 0.4000 | 0.6419 |
| | | | Min | 0.9000 | 0.8000 | 0.7667 | **0.7250** | **0.7000** | **0.8000** | 0.5500 | 0.5000 | **0.4750** | **0.4800** | **0.6877** |
| | | | Mean | **1.0000** | **0.9000** | **0.8000** | 0.6000 | 0.6200 | 0.7000 | **0.7500** | **0.5333** | 0.4500 | 0.4000 | 0.6738 |
| | | | D-S | 0.9000 | 0.6500 | 0.6667 | 0.6000 | 0.6000 | 0.7000 | 0.4000 | 0.4333 | 0.4000 | 0.3800 | 0.6408 |
| | | | Yager | 0.9000 | 0.6500 | 0.6667 | 0.6000 | 0.6000 | 0.7000 | 0.4000 | 0.4333 | 0.4000 | 0.3800 | 0.6408 |
| | | | CA | **1.0000** | **0.9000** | **0.8000** | 0.6000 | 0.6400 | 0.7000 | **0.7500** | **0.5333** | 0.4500 | 0.4000 | 0.6753 |
| Factoid Q&A | 5 | Baseline | Google | 0.3897 | 0.4717 | 0.5088 | 0.5407 | 0.5700 | 0.0449 | 0.0872 | 0.1501 | 0.2719 | 0.3782 | 0.5505 |
| | | | Yahoo! | 0.7701 | 0.6978 | 0.6548 | 0.6256 | 0.6018 | 0.2707 | 0.2844 | 0.3419 | 0.3706 | 0.3921 | 0.5850 |
| | | | Bing | 0.8530 | 0.8025 | 0.7579 | 0.7261 | 0.7027 | 0.2544 | 0.2800 | 0.3257 | 0.3703 | 0.4042 | 0.5799 |
| | | Comb. Rule | Max | 0.7538 | 0.7124 | 0.7159 | 0.7007 | 0.6830 | 0.1599 | 0.2287 | 0.2900 | 0.3852 | 0.4312 | 0.5781 |
| | | | Min | 0.7404 | 0.6689 | 0.6136 | 0.5828 | 0.5582 | 0.3302 | 0.2973 | 0.3399 | 0.3638 | 0.3858 | 0.6063 |
| | | | Mean | **0.8670** | 0.8445 | 0.7950 | 0.7636 | 0.7366 | **0.3349** | **0.3474** | 0.3987 | 0.4586 | 0.5056 | 0.6218 |
| | | | D-S | 0.6989 | 0.5674 | 0.5441 | 0.5373 | 0.5261 | 0.3191 | 0.2681 | 0.2818 | 0.3299 | 0.3387 | 0.6001 |
| | | | Yager | 0.6989 | 0.5674 | 0.5441 | 0.5373 | 0.5261 | 0.3191 | 0.2681 | 0.2818 | 0.3299 | 0.3387 | 0.6001 |
| | | | CA | 0.8652 | **0.8448** | **0.7966** | **0.7649** | **0.7390** | 0.3372 | 0.3468 | **0.3993** | **0.4589** | **0.5069** | **0.6221** |
| | 10 | Baseline | Google | 0.2742 | 0.4495 | 0.5270 | 0.5665 | 0.5610 | 0.0088 | 0.0630 | 0.1725 | 0.2655 | 0.3484 | 0.5433 |
| | | | Yahoo! | 0.7742 | 0.7115 | 0.6636 | 0.6294 | 0.6057 | 0.2707 | 0.2768 | 0.3359 | 0.3617 | 0.3908 | 0.5831 |
| | | | Bing | 0.8337 | 0.7812 | 0.7495 | 0.7167 | 0.6860 | 0.2935 | 0.2973 | 0.3086 | 0.3547 | 0.3876 | 0.5807 |
| | | Comb. Rule | Max | 0.7497 | 0.7211 | 0.7001 | 0.6944 | 0.6844 | 0.2602 | 0.2447 | 0.3145 | 0.3765 | 0.4350 | 0.5905 |
| | | | Min | 0.7334 | 0.6844 | 0.6328 | 0.6098 | 0.5944 | **0.3279** | 0.2940 | 0.3038 | 0.3508 | 0.3895 | 0.5983 |
| | | | Mean | 0.8390 | 0.8238 | **0.7962** | **0.7754** | 0.7540 | 0.3273 | **0.3296** | 0.4082 | 0.4545 | 0.5034 | **0.6190** |
| | | | D-S | 0.6931 | 0.5522 | 0.5511 | 0.5395 | 0.5300 | 0.3162 | 0.2620 | 0.2832 | 0.3267 | 0.3378 | 0.5916 |
| | | | Yager | 0.6931 | 0.5522 | 0.5511 | 0.5395 | 0.5300 | 0.3162 | 0.2620 | 0.2832 | 0.3267 | 0.3378 | 0.5916 |
| | | | CA | **0.8396** | **0.8258** | 0.7946 | **0.7754** | **0.7544** | **0.3279** | 0.3288 | **0.4084** | **0.4549** | **0.5040** | **0.6190** |

on the match between the content of the site and the question. Our main reported findings have significant implications for the design of social Q&A systems.

In our observations, we also find that the content preference of social media sites, whatever they are, are slowly changing over time. It is easy to understand since the users' interests, social sites' characteristics, and the outside world keep changing. However, time-varying content poses a great challenge to dynamically capture and understanding the nature of social sites, and this also is our future work.

## ACKNOWLEDGMENT

This work was sponsored, in part, by a research fund from National Security Innovations (NSI), the National Natural Science Foundation of China (61001178), the Excellent Talents Foundation of Beijing, the Importation and Development of High-Caliber Talents Project of Beijing Municipal Institutions (CITTCD201404052), and the Basic Research Program of Beijing University of Technology.

## REFERENCES

[1] R. Zafarani and H. Liu, "Users joining multiple sites: Distributions and patterns," in *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*, (Ann Arbor, MI), pp. 635–638, AAAI Press, June 2014.

[2] M. Efron and M. Winget, "Questions are content: A taxonomy of questions in a microblogging environment," in *Proceedings of the American Society for Information Science and Technology*, (Austin, Texas), pp. 1–10, AAAI Press, August 2010.

[3] A. Paul, L. Hong, and H. Chi, "Is twitter a good place for asking questions?: A characterization study," in *Proceedings of the International Conference on Weblogs and Social Media*, (Barcelona, Spain), pp. 17–21, AAAI Press, July 2011.

[4] R. Morris, J. Teevan, and K. Panovich, "What do people ask their social networks, and why?: A survey study of status message q&a behavior," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, (Atlanta, GA), pp. 1739–1748, ACM, April 2010.

[5] L. M. Aiello, R. Schifanella, and B. State, "Reading the source code of social ties," in *Proceedings of the 2014 ACM Conference on Web Science*, WebSci'14, (New York, NY, USA), pp. 139–148, ACM, 2014.

[6] P. Dempster, "Upper and lower probabilities induced by a multivalued mapping," *The annals of mathematical statistics*, vol. 38(2), pp. 325–339, 1967.

[7] R. R. Yager, "On the dempster-shafer framework and new combination rules," *Information Sciences*, vol. 41(2), pp. 93–137, 1987.

[8] A. Zadeh, "A simple view of the dempster-shafer theory of evidence and its implication for the rule of combination," *AI magazine*, vol. 7(2), pp. 85–85, 1986.

[9] S. Beitzel, E. Jensen, A. Chowdhury, D. Grossman, and O. Frieder, "Hourly analysis of a very large topically categorized web query log," in *Proceedings of the 27th annual international ACM SIGIR conference*, (Sheffield, UK), pp. 321–328, ACM, July 2004.

[10] J. Grajales III, S. Sheps, and K. Ho, "Social media: A review and tutorial of applications in medicine and health care," *Journal of Medical Internet Research*, vol. 16(2), pp. e13–e13, 2014.

[11] N. Smith, M. Heilman, and R. Hwa, "Question generation as a competitive undergraduate course project," in *Proceedings of the NSF Workshop on the Question Generation Shared Task and Evaluation Challenge*, (Arlington, VA), September 2008.

[12] Z. Yang, J. Lei, K. Fan, and Y. Lai, "Keyword extraction by entropy difference between the intrinsic and extrinsic mode," *Physica A: Statistical Mechanics and its Applications*, vol. 392(19), pp. 4523–4531, 2013.