Social Answer: A System for Finding Appropriate Sites for Questions in Social Media

Harsh Dani*, Fred Morstatter*, Xia Hu ‡, Zhen Yang† and Huan Liu*
*Computer Science and Engineering, Arizona State University, Tempe, AZ, USA
†Computer Science and Engineering, Texas A&M University, USA
‡College of Computer Science, Beijing University of Technology, Beijing, China
*{harsh.dani, fred.morstatter, huan.liu}@asu.edu, ‡hu@cse.tamu.edu, †yangzhen@bjut.edu.cn

***Abstract*—Social networking or social media sites are involved in our daily life. With the increasing popularity of social media sites like Twitter and Facebook, people are using their social network to find answers to their questions. Not every social media site can answer a user's question. Different social media sites have different strength such as the StackOverflow community has specific interest in programming and software development, whereas the TripAdvisor community has specific interest in topic of Travel. Hence, we need a framework to identify social media sites, which can best answer user's questions. In this paper, we demonstrate system named `Social Answer` which can rank social media sites for user's query based on the content similarity of social media site and user's query using ensemble of search engine query results.**

***Keywords*-Social Media; Q&A; Conflict Allocation.**

I. Introduction

Social media or social networking sites are involved in our daily life. Social networks open up possibilities for discovering new information, sharing ideas and interacting with others. With the increasing popularity of social networking sites like Twitter and Facebook, people are using their social networks to find answers to their questions [2]. Social media sites help users to gather information, advice and expertise [3]. Nowadays, social media sites are often used for Question & Answer purposes and there are many advantages of asking questions on social media sites.

A previous study done in Morris et al. compared information seeking using search engines and social networks [4]. The authors concluded that when users need opinion-type answers, they prefer individuals who know them. Users' friends know additional context about them and can provide with tailored answers. Social networks are particularly useful for asking subjective questions. People consider results from their social network highly trustworthy. Also, there are many social media sites like Xing[1], which is a platform for professional networking and Raptr[2], which is a community for PC and console gaming, are some of social media sites, which users are not aware of. Asking question on all social media sites is time consuming, infeasible and not every question is answered on all social media platforms due to their content restrictions to some specific topics.

Different social media sites have different strengths and users participating in that social media platform are interested in certain topics. For example, the StackOverflow community has specific interest in programming and software development, where the TripAdvisor community has specific interest in topic of Travel. For example, the question "Where can I find a good Mexican restaurant near me?" is better suited for Yelp, instead of StackOverflow. The goal of `Social Answer` is to identify the correct social media site for a question where, other users participating in that social media site possess knowledge to potentially answer that question satisfactorily. Finding the appropriate social media sites for a user's query can be challenging at times.

Users' queries are generally short in length, have implied meaning and are noisy. As content of social media sites are changing quickly and new content is generated everyday, indexing social media sites frequently is not feasible. We need a method to dynamically retrieve the content of social media sites. We retrieve the content of social media sites using a search engine, as search engines update and index new pages continuously. However, depending on single search engine is not reliable due to their unknown ranking criterion. Therefore, we need a method to aggregate the search engine results from different search engines for a given social media site to compute the measure of similarity of social media site to a user's query.

In this paper, we demonstrate a system `Social Answer` that can rank social media sites for question using a "conflict allocation" algorithm [1] which is an ensemble of search engine query results.

II. Technical Specifications: System Overview

In this section, we formally introduce the problem. We first introduce the notations used. $q = (q_1, q_2, ..., q_n)$ denotes the question asked by user where each $q_i \in q$ denotes the words in the question. $k = (q_1, ..., q_k)$ denotes the keywords in the question. After all keywords have been extracted from question, we expand each keyword $q_k \in k$ via Wikipedia. All returned Wikipedia articles are indexed and their word

[1] https://www.xing.com

[2] http://www.raptr.com

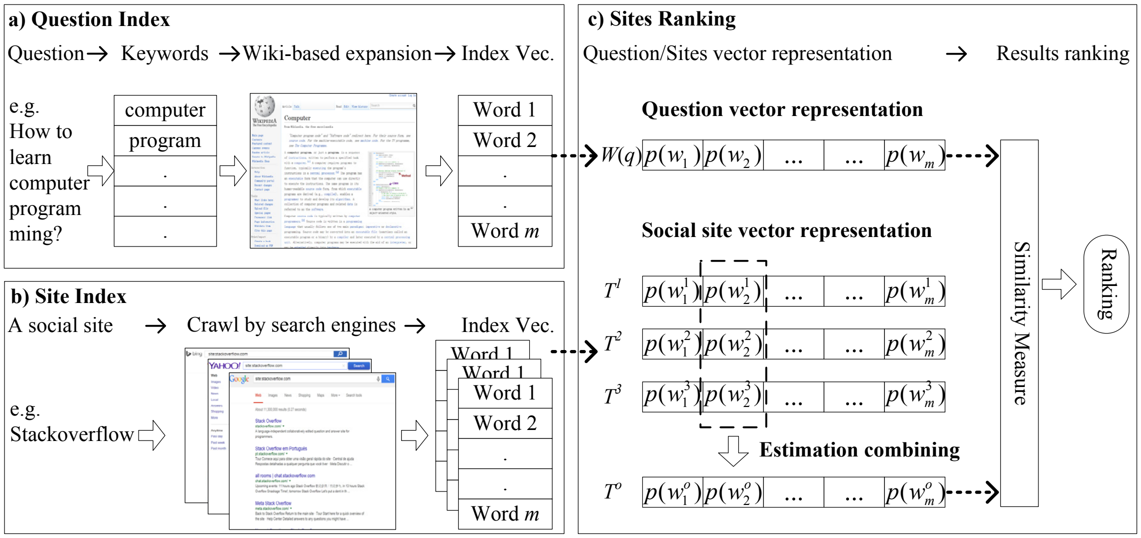

a) Question Index
Question→ Keywords →Wiki-based expansion→ Index Vec.
e.g. How to learn computer programming?
computer
program
Word 1
Word 2
Word m
b) Site Index
A social site → Crawl by search engines → Index Vec.
e.g. Stackoverflow
Word 1
Word 2
Word m
c) Sites Ranking
Question/Sites vector representation → Results ranking
Question vector representation
W(q) p(w_1) p(w_2) ... ... p(w_m)
Social site vector representation
T^1 p(w_1^1) p(w_2^1) ... ... p(w_m^1)
T^2 p(w_1^2) p(w_2^2) ... ... p(w_m^2)
T^3 p(w_1^3) p(w_2^3) ... ... p(w_m^3)
Estimation combining
T^o p(w_1^o) p(w_2^o) ... ... p(w_m^o)
Similarity Measure
Ranking

Figure 1. Block diagram of `Social Answer` obtained from [1]. a) Shows keyword extraction and question expansion using Wikipedia. First, the keywords are extracted from question and extracted keywords are then expanded using Wikipedia. Index vector of the content obtained from Wikipedia is created. b) Shows search engine crawling. Each social media site's content is obtained by different search engines with domain restriction to that specific social media site. c) Shows the combining evidence obtained from different search engine and finding similarity between question vector and social media sites vector.

frequency vector is denoted by $W(q)$. We use S to denote set of social media sites, where each $s_i \in S$ is a candidate social media site. We use g to denote the set of search engines where each $g_j \in g$ is a individual search engine such as Google or Yahoo. We use $T(s_i, g_j, n)$ to denote the word frequency vector of the top-n pages from search engine g_j within site s_i's domain. We denote $P(w_j^i)$ as the frequency estimate of word j by search engine i. $P(w_j^\circ)$ denotes the optimal estimate of word j by M different search engines. We use $\|.\|$ to denote the l_2 norm of vector.

The formal problem statement can be given as: Given a question q and set of social media sites S where each $s_i \in S$ is a candidate social media site, the goal of `Social Answer` is to find the set of ranked social media sites S', where each candidate social media site $s'_i \in S'$ is ranked according to its probability of answering question q.

The questions asked by users are generally short, implicit and noisy [1]. To tackle this challenge, we identify keywords k in the question. Then we take each keyword $q_k \in k$ and expand their Wikipedia articles using Wikipedia's Python API interface and index all returned Wikipedia articles as word frequency vector $W(q)$. The expanded Wikipedia articles are appended together and sent to system as a question, instead of user's query, as Wikipedia articles can model the user's intent better than short question [1].

There are many tools and public API interfaces available to crawl the content from different social media sites. However, there is not an effective way to retrieve the most expressive content of social media sites[1]. Hence, social media sites were explored through the search engines. For each candidate social media site $s_i \in S$, we crawled its top-n pages using search engine g_j and created word frequency vector $T(s_i, g_j, n)$. For question q and search engine g_j, we can rank each site $s_i \in S$ using cosine similarity and can be formulated as [1]:

$$D(s_i, q) = \frac{< T(s_i, g_j, n), W(q) >}{\|T(s_i, g_j, n)\| \|W(q)\|} \tag{1}$$

Due to search engine's unknown ranking criterion, using results from a single search engine is not reliable and consistent. It poses a challenge of how to combine the results from different search engines, since the results from different search engines might be similar or highly conflicting.

In `Social-Answer`, we have used the *Conflict allocation* [1] ensemble which can be formulated as:

$$p(w_j^\circ) = \sum_{\cap_i w_j^i = w_j} \prod_i p(w_j^i) + \tag{2}$$

$$q(w_j)(1 - \sum_j \sum_{\cap_i w_j^i = w_j} \prod_i p(w_j^i))$$

Where $q(w_j)$ is defined as:

$$q(w_j) = \frac{\sum_i p(w_j^i)}{\sum_i \sum_j p(w_j^i)} \tag{3}$$

In Equation 2, $\sum_{\cap_i w_j^i = w_j} \prod_i p(w_j^i)$ denotes sharing measure and combines the probability from different sources in the same manner as Yager [5]. $(1 - \sum_j \sum_{\cap_i w_j^i = w_j} \prod_i p(w_j^i))$ denotes the conflicting probability between different search engines. The conflicting belief is assigned according to weight $q(w_j)$ [1].

Table I
CANDIDATE SOCIAL MEDIA SITES

Category	Site (Alexa rank)
Blogs	Wordpress(26), Blogger(53)
Microblogs	Twitter(7)
Social Networks	Facebook(2)
Professional Networks	Linkedin(12), Xing(908), Viadeo(1838)
Media Content Sharing	Youtube(3), Pinterest(25), Instagram(30)
	Tumblr(39), Imgur(49), Flickr(103)
Collaborative Knowledge Base	Wikipedia(6)
Social filter	Reddit(50), Yelp(125)
Collaborative Q&A	Answers(200), StackOverflow(45)

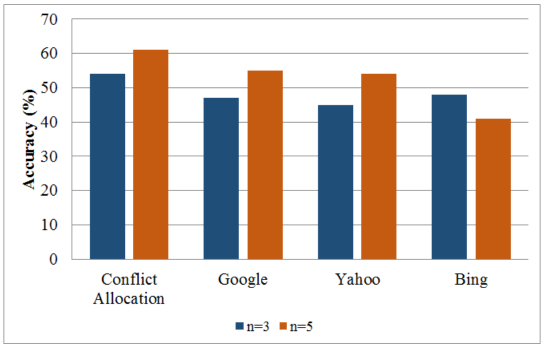


Figure 2. Performance comparison of demonstrated system with baselines a) Searching on Google b) Searching on Yahoo c) Searching on Bing. Y-axis shows accuracy score.

III. System Demonstration

In this section we demonstrate `Social-Answer` which has three stages: 1) Search Engine Crawling 2) Query expansion using Wikipedia 3) Question and search engine vector similarity.

A. Search Engine Crawling

Firstly, we crawled the content of 25 social media sites listed in Table I. We use 3 different search engines: Google, Yahoo, and Bing. We took each social media site and query empty string with domain restriction to that social media site (e.g site:www.yelp.com). We took top 10 pages from each of the search engine. We removed stopwords, numbers and URLs from the content and performed stemming on each of these words and indexed them. We used the Python programming language to implement the crawling module and to perform preprocessing steps.

B. Query Expansion using Wikipedia

In this stage, we first extracted keywords from the question. We consider all nouns present in the question as keywords. The question asked by users are generally short, noisy and have implicit meaning. Also, these questions can lead to very sparse representations. To tackle this problem, we used Wikipedia to expand keywords. We obtained the Wikipedia article of each keyword using Wikipedia's Python (API). We also removed stopwords, numbers, URLs from the text obtained from Wikipedia and performed stemming. The content obtained from the Wikipedia article is then sent to the system as a question. We used the Python programming language to perform keyword extraction Wikipedia expansion and to perform preprocessing. The web application of `Social Answer` was developed using Flask[3], which is a python web application development framework.

C. Question and Search Engine Vector Similarity

In this stage, we measured the similarity between the content obtained from Wikipedia and content indexed from search engine in order to rank social media sites for specific questions. We first obtained a frequency estimate of each word in the corpus of content indexed from search engine crawling and content obtained from Wikipedia. Then we combined estimates from different search engines using conflict allocation algorithm described in Section II. Then we use the cosine similarity measure defined in Equation 1 to obtain the similarity score. The top 5 social media sites having the highest similarity score are reported back to user that are most likely to answers user's question.

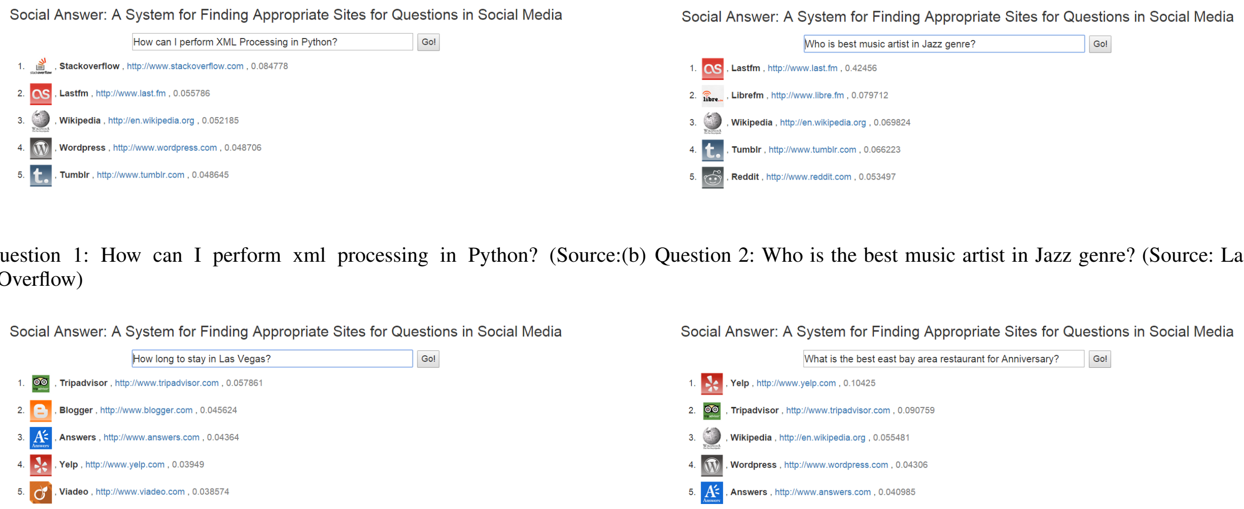
IV. Evaluation

We performed extensive evaluation of the `Social Answer`. With this experiment, we aim to answer following question: "How effective is the proposed framework compared to its baselines?"

We consider three baselines: 1) Searching on Google 2) Searching on Bing 3) Searching on Yahoo. To perform the experiments, we collected 100 questions that are already answered, from candidate social media sites listed in Table I and consider those social media sites as the ground truth. We define *accuracy score* as: Whether the ground truth social media site is in the top n social media sites returned by our system and baselines. We consider two values of n, $n = 3$ and $n = 5$.

In experiments, we use the same 25 social media sites as [1]. As suggested by many researchers in their previous work [6], the social media sites were divided into 10 categories: Blogs, Microblogs, Professional networks, Social networks, Collaborative knowledge base, Media content sharing, Collaborative question & answering, Social filtering, Instant messaging and Virtual social and game worlds. In our experiments, we selected 17 social media sites from above categories from top 200 social media site listed on Alexa.com and 8 other well known social media sites [1]. Due to privacy issues, instant messaging and virtual social & game world categories were not investigated. Table I shows some of the selected social media site from each category and their corresponding rank on Alexa [1].

[3] http://flask.pocoo.org/



Social Answer: A System for Finding Appropriate Sites for Questions in Social Media

How can I perform XML Processing in Python?

1. , **Stackoverflow** , http://www.stackoverflow.com , 0.084778
2. , **Lastfm** , http://www.last.fm , 0.055786
3. , **Wikipedia** , http://en.wikipedia.org , 0.052185
4. , **Wordpress** , http://www.wordpress.com , 0.048706
5. , **Tumblr** , http://www.tumblr.com , 0.048645

Social Answer: A System for Finding Appropriate Sites for Questions in Social Media

Who is best music artist in Jazz genre?

1. , **Lastfm** , http://www.last.fm , 0.42456
2. , **Librefm** , http://www.libre.fm , 0.079712
3. , **Wikipedia** , http://en.wikipedia.org , 0.069824
4. , **Tumblr** , http://www.tumblr.com , 0.066223
5. , **Reddit** , http://www.reddit.com , 0.053497

(a) Question 1: How can I perform xml processing in Python? (Source: StackOverflow)

(b) Question 2: Who is the best music artist in Jazz genre? (Source: Last.fm)

Social Answer: A System for Finding Appropriate Sites for Questions in Social Media

How long to stay in Las Vegas?

1. , **Tripadvisor** , http://www.tripadvisor.com , 0.057861
2. , **Blogger** , http://www.blogger.com , 0.045624
3. , **Answers** , http://www.answers.com , 0.04364
4. , **Yelp** , http://www.yelp.com , 0.03949
5. , **Viadeo** , http://www.viadeo.com , 0.038574

Social Answer: A System for Finding Appropriate Sites for Questions in Social Media

What is the best east bay area restaurant for Anniversary?

1. , **Yelp** , http://www.yelp.com , 0.10425
2. , **Tripadvisor** , http://www.tripadvisor.com , 0.090759
3. , **Wikipedia** , http://en.wikipedia.org , 0.055481
4. , **Wordpress** , http://www.wordpress.com , 0.04306
5. , **Answers** , http://www.answers.com , 0.040985

(c) Question 3: How long to stay in Las Vegas? (Source: TripAdvisor)

(d) Question 4: What is good mexican restaurant near me? (Source: Yelp)

Figure 3. Screenshots of `Social Answer`

As shown in the Figure 2, `Social Answer` shows compelling results and beats its baselines by 6%. Also, Figure 3 shows the screenshots of `Social Answer`.

V. Related Work

Many researchers have worked to understand information seeking in social media. Previous research has found that information seeking in social media is strongly tied to users' natural interactions not similar to information seeking in traditional Information Retrieval environment [7].

Paul et al. [2] conducted a study of Q&A behavior on Twitter and found that most questions were rhetorical and factual ones. Teevan et al. [8] investigated the top category of questions sent to Aardvark, which is a social search engine and found that questions were experience-oriented, recommendation, local and advice queries. Morris et. al [9] did an extensive survey on what people usually ask their social network and found that most typical questions that users asked on social networks were opinion based.

VI. Conclusion

We demonstrated an interactive system `Social Answer` which ranks social media sites to users' question. We extract keywords from the question and expand them using Wikipedia also we retrieve the most expressive content of social media sites using ensemble of Google, Yahoo ,and Bing. We use "Conflict allocation" ensemble to combine evidence obtained from different search engines and use cosine similarity measure to rank the social media sites for questions. `Social Answer` yields compelling results compared to its baselines and beats them by 6%.

References

[1] Z. Yang, I. Jones, X. Hu, and H. Liu, "Finding the right social media sites for questions," *ASONAM*, 2015.

[2] S. A. Paul, L. Hong, and E. H. Chi, "Is twitter a good place for asking questions? a characterization study." in *ICWSM*, 2011.

[3] G. Y. Jeon and S. Y. Rieh, "The value of social search: Seeking collective personal experience in social q&a," *Proceedings of the American Society for Information Science and Technology*, vol. 50, no. 1, pp. 1–10, 2013.

[4] M. R. Morris, J. Teevan, and K. Panovich, "A comparison of information seeking using search engines and social networks." *ICWSM*, vol. 10, pp. 23–26, 2010.

[5] R. R. Yager, "On the dempster-shafer framework and new combination rules," *Information sciences*, vol. 41, no. 2, pp. 93–137, 1987.

[6] J. H. Kietzmann, K. Hermkens, I. P. McCarthy, and B. S. Silvestre, "Social media? get serious! understanding the functional building blocks of social media," *Business horizons*, vol. 54, no. 3, pp. 241–251, 2011.

[7] J. Yang, M. R. Morris, J. Teevan, L. A. Adamic, and M. S. Ackerman, "Culture matters: A survey study of social q&a behavior." *ICWSM*, vol. 11, pp. 409–416, 2011.

[8] J. Teevan, D. Ramage, and M. R. Morris, "# twittersearch: a comparison of microblog search and web search," in *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM, 2011, pp. 35–44.

[9] M. 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*. ACM, 2010, pp. 1739–1748.