Answering Twitter Questions: a Model for Recommending Answerers through Social Collaboration

Laure Soulier, Lynda Tamine, Gia-Hung Nguyen

Answering Twitter Questions: a Model for Recommending Answerers through Social Collaboration

Laure Soulier
Sorbonne Universités - UPMC
Univ Paris 06, CNRS LIP6
UMR 7606
75005 Paris, France
laure.soulier@lip6.fr

Lynda Tamine
Université de Toulouse,
UPS - IRIT
118 route de Narbonne
31062 Toulouse, France
tamine@irit.fr

Gia-Hung Nguyen
Université de Toulouse,
UPS - IRIT
118 route de Narbonne
31062 Toulouse, France
gia-hung.nguyen@irit.fr

ABSTRACT

In this paper, we specifically consider the challenging task of solving a question posted on Twitter. The latter generally remains unanswered and most of the replies, if any, are only from members of the questioner's neighborhood. As outlined in previous work related to community Q&A, we believe that question-answering is a collaborative process and that the relevant answer to a question post is an aggregation of answer nuggets posted by a group of relevant users. Thus, the problem of identifying the relevant answer turns into the problem of identifying the right group of users who would provide useful answers and would possibly be willing to collaborate together in the long-term. Accordingly, we present a novel method, called CRAQ, that is built on the collaboration paradigm and formulated as a group entropy optimization problem. To optimize the quality of the group, an information gain measure is used to select the most likely "informative" users according to topical and collaboration likelihood predictive features. Crowd-based experiments performed on two crisis-related Twitter datasets demonstrate the effectiveness of our collaborative-based answering approach.

Keywords

Social information retrieval, Collaborative group recommendation, Social Network Question-Answering

1. INTRODUCTION

A recent Pew Internet survey[1] published in August 2015 indicates that Facebook and Twitter are the most prominent social media services, with more than 72% and 23% of US adults utilizing the services, respectively. Although social platforms were designed to create social connections, they have emerged as tractable spaces for information seeking through the posing of questions [29]. A survey on question-asking practices on Twitter and Facebook reveals that over 50% of users ask questions on social networks [40]. The main reason users prefer a social platform to a search engine is that they trust their social network's replies and specifically seek subjective replies (e.g., opinions/advice) rather than objective information provided by search engines.

However, there are several issues related to asking questions on social media. One issue is that the majority of questions do not receive a reply, while a minority receives a high number of responses [21, 32]. Another issue is that even though questioners use manual question-oriented hashtags (e.g., #lazyweb, #twoogle) to bypass the local neighborhood [19], the answers are mostly provided by members of the immediate follower network, characterizing behavior patterns known as friendsourcing [20]. These issues give rise to other drawbacks. First, users are uncomfortable with posing private questions (e.g., religious or political) to their social neighbors [29]. Second, recent research has highlighted that questioners perceive that friendsourcing has a social cost (e.g., spent time and deployed effort) that weakens the potential of social question-answering [20, 30].

Hence, an effective recommendation for skilled answerers responding to questions posted online for a wide audience is highly desirable. In this perspective, a common approach early developed in community Q&A services [24, 22] and in other popular social networks, such as Facebook and Twitter [17, 25], consists in routing the questions to a list of the top-k appropriate users. The users' appropriateness is generally estimated using a set of features from the questions, the users themselves and their relations with the questioner. However, collaboration has been acknowledged as a valuable method for completing information search tasks within [28, 9] (e.g., the DARPA challenge [39]) and outside of social media spaces [12]. Therefore, another approach that we believe to be valuable, is to consider the question-answering task as a collaborative process involving a group of socially authoritative users with complementary skills. Such users would allow for the gathering of diverse pieces of information related to the question, thereby reinforcing the likelihood that their answers are relevant as a whole. Accordingly, we address the problem of tweet question solving by turning it into the problem of collaborative group building which is the core contribution of this paper. More specifically, based on the information gain theory, our approach favors the selection of a collaborative group of answerers rather than the ranking of answerers. Our long-term goal is to favor explicit social collaboration between the group members including the questioner.

[1] http://www.pewinternet.org/2015/08/19/the-demographics-of-social-media-users/

Based on this driving idea, we propose the CRAQ-Collaborator Recommendation method for Answering Twitter Questions. Given a posted question q and topically relevant tweets T posted by an initial group U of users, the general schema supporting the CRAQ relies on an evolving process of group building that discards users who supply the lowest information gain at each step, leading to the maximization of the group information entropy [33].
To summarize, this paper makes the following contributions:

• First, we introduce a novel algorithm recommending a collaborative group of users to tweeted questions so as a cohesive answer could be inferred from the aggregation of their topically relevant tweets. To the best of our knowledge, this is the first effort to recommend a social collaborative group of users aiming at maximizing the overall relevance of answers to a tweeted question.

• Second, we build a predictive model to characterize the collaboration likelihood between pairwise users who are collectively able to solve target tweeted questions. The model is built on both the user's authority and the complementarity of the content of their tweets.

• Third, we perform an empirical crowdsourced-based evaluation which shows that collaborative-based tweet question answering is effective.

In what follows, Section 2 reviews the relevant prior work. Section 3 presents the problem and provides an overview of the CRAQ. Section 4 details the predictive model of collaboration and the group recommendation algorithm. Section 5 details the experimental setup, and the results are presented in Section 6. Last, we conclude in Section 7.

2. RELATED WORK

Social Network Question Answering (SNQ&A).

Unlike in community Q&A [24, 22, 9], users who post questions on social network sites are engaged in information seeking with a wide range of users from their social cluster or even strangers. Numerous research studies have investigated the characteristics of questions addressed to social networks, including the topics and the motivations behind the questions [29, 32, 21]. Through a survey of Facebook and Twitter users, Morris et al. [29] showed that most users post questions to obtain subjective (opinion-based and engaged) answers related to a wide range of interests, including technology, entertainment, home, and family. Considering the benefits of engaging in social collaboration while performing a search [17, 28, 15], Horowitz et al. [17] designed the *Aardvark* social search engine built on the "village paradigm". The authors' approach consists in routing the question to different people using a retrieval task aiming at identifying the appropriate person to answer the question. Formally, the authors proposed a probabilistic model based on (1) users' expertise, (2) the connectedness of the recommended user and the questioner, and (3) the availability of the candidate answerer. Similarly, Hecht et al. [15] presented a socially embedded search engine, called *Search Buddies* and collocated with Facebook. This system includes two main functionalities that are proposed to answer questions: recommending relevant posts or messages that are likely to answer a question (*Investigator API*) and connecting to people who may have the answer (*Butterfly API*). One limitation of these two previous studies is that they tracked information and users within the questioners' neighborhood. This method might be restrictive and hinder the likelihood of obtaining replies [21, 32]. Unlikely, one recent work [25] which is most related to this paper attempt to automatically identify appropriate users, even strangers, to answer tweeted questions. The users' appropriateness was estimated based on their willingness to answer the question and readiness to answer the question shortly after it was submitted. Statistical models were used to predict the probability that a user would provide an answer to the question. Second, based on the likelihood that a user answers the question, a ranked list of users is retrieved using a classification algorithm identifying users as responders or non-responders. Experiments based on three datasets collected via a human operator who presented questions to target users on Twitter showed that the recommendation algorithm improves the response rate compared with random users (by 59%) and a simple binary classification of target users (by 8%). In contrast to these authors' work, we aim to recommend a collaborative group of users viewed here as authors of diverse and complementary pieces of answers with a maximal expected relevance to a given question. To the best of our knowledge, it is the first attempt in the context of SNQ&A for providing a collaborative-based answer to tweet questions. Beyond the goal of providing the relevant answer to a tweet question, our method can be applied to recommend users and then favor long-term collaboration among questioners and answerers.

Group selection.

Group selection [2, 13] is one of the key underlying issues addressed in a long-standing and multidisciplinary study on collaboration [3]. Regardless of the contexts in which it is embedded, explicit (implicit) collaboration typically refers to individuals' action of intentionally (unintentionally) working together to complete a shared task. For instance, Augustin-Blas et al. [2] proposed a genetic algorithm-based model for group selection. The objective function of the algorithm consists of maximizing the shared knowledge resources within the selected candidate teams. This approach has been shown to be effective in solving the issue of teaching group formation. The study conducted by González-Ibáñez et al. [13] focused on estimating the costs and benefits of pseudo-collaboration between users with similar information search tasks. In pseudo-collaboration, users performing a search task are given recommendations based on the past results obtained by other users who performed similar tasks. An experimental study revealed that the efficiency of group selection based on pseudo-collaboration was greater than that the one of the group selection including users who intentionally shared a search task through explicit collaboration. More recent studies specifically addressed group selection within social media platforms [8, 6]. Castilho et al. [8] investigated whether the social behavior of candidate collaborators is an important factor in their choices when forming a collaborative task team. The authors performed an analysis using data related to the Facebook profiles of students in the same class. The students were asked to organize themselves into groups to perform a simulated task. The study results clearly showed that beyond proficiency, strength of friendship and popularity were determinant for the group selection. As further confirmation of this phenomenon, the most skilled users were not always preferred. From another perspective,

Cao et al. [6] exploited URL-sharing information to design a classifier that was able to distinguish between organized and organic groups on Twitter. Their experimental study showed that organic groups' shared topics were more focused and sharing behaviors were more similar compared with those of organized groups. Another line of research examined group selection for task allocation on crowdsourcing platforms [23, 34, 1]. Li et al. [23] proposed an effective predictive model for identifying accurate working groups based on workers' attributes relevant to the task characteristics. Rahman et al. [34] noted that beyond worker skills and wages, worker-worker affinity is a key element to consider optimizing collaboration effectiveness. Abraham et al [1] focused on determining the optimal size of a group of workers allowing to achieve a good balance between the cost and quality outcome of a human intelligence task such as label quality assessment. The authors proposed an adaptive stopping rule which decides during the process of worker hiring whether the optimal group size of crowdworkers has been reached. The rule mainly relies on the level of task difficulty and workers' skills to achieve the task.

In summary, previous work unveils interesting findings among which group performance is affected by three main factors captured at both individual and group levels: (1) members' knowledge and skills with regard to the task requirements, (2) members' authority within the social network measuring the users' expertise and trust towards the task, and (3) group size allowing a good quality-cost tradeoff to achieve the collaborative task.

3. PROBLEM FORMULATION AND METHOD OVERVIEW

Based on the literature review highlighting the challenge of solving an IR task in the context of SNQ&A [29] and the potential of collaboration within social networks [17, 28] as well as IR tasks [37, 13], we address the problem of recommending social collaborators for a tweeted question. To do so, we design a task that identifies a group of complementary answerers who could provide the questioner with a cohesive and relevant answer. Taking into account the long-term perspective of gathering users who are willing to collaborate, we carefully consider the problem of group size optimization in terms of balance between the benefits of collaboration and the cognitive cost underlying crowdsourcing [20, 1]. Therefore, for a tweeted question q, the main issue consists in identifying the smallest collaborative group $g \subset U$ in which each user $u_j \in g$ can contribute to providing a relevant and cohesive answer to the given question q. Accordingly, we believe that the intuition underlying the level at which user u_j contributes given collaborative group g could be assimilated to an information gain-based metric noted as $IG(g, u_j)$. Instead of maximizing the average of information gain over users (which could be seen as a NP-hard problem as mentioned in [34]), we propose the CRAQ in which, given a posted question q and an initial group U of users, we build the collaborative group $g \subset U$ iteratively by discarding users providing the lowest information gain, leading to the maximization of the group entropy [33]. We provide a general overview of our two-phase CRAQ general methodology in Figure 1. In phase A (Section 4.2), we build a predictive model offline that can predict the likelihood of users' pairwise collaboration. To attain this goal, we rely on

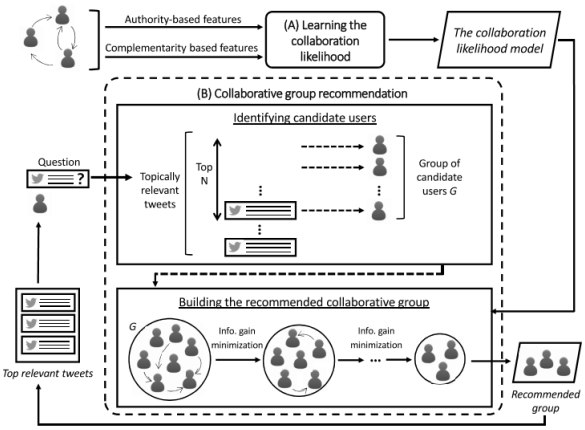

Figure 1: Overview of the CRAQ-based methodology

the finding that group performance is affected by two main factors [23, 34]: (1) members' knowledge and skills relevant to the task requirements, and (2) members' authority within the social network. With this in mind, we develop a predictive model of a collaborative group based on group members complementarity and authority. In phase B (Section 4.3), given a tweeted question, we first build a group of candidate users who posted topically relevant tweets. Then, considering the predictive model of collaboration built in phase A, we iteratively discard *non-informative* users and build a smaller group of answerers to recommend to the questioner by ensuring that each user in the group contributes to the collaborative response. This contribution is estimated using the group-based information gain metric. Finally, top K tweets posted from each of the candidate answerers are returned as a whole answer to the user's question.

4. METHOD

4.1 Preliminaries

Before detailing our method for recommending a group of social collaborators, we introduce the underlying key concept related to the group-based information gain, noted $IG(g, u_k)$. The latter estimates each user's $u_k \in g$ contribution to group g with respect to question q and their complementarity. The metric is estimated as the difference between the group entropy $H(g)$ and conditional entropy $H(g|u_k)$:

$$IG(g, u_k) = H(g) - H(g|u_k) \quad (1)$$

These two entropy-based metrics denote the two assumptions related to the topical relevance of users' tweets and users' skills within the collaborative group:

• Group entropy $H(g)$ measures the amount of information provided by all users u_j belonging to group g given query q (Eq. 2). Because group entropy is mostly related to the topic of the question, we estimate $p(u_j)$ according to a user-query similarity-based probability $p(u_j|q)$ (Eq. 3):

$$H(g) = -\sum_{u_j \in g} P(u_j) \cdot log(P(u_j)) \quad (2)$$

$$H(g) \propto -\sum_{u_j \in g} P(u_j|q) \cdot log(P(u_j|q)) \quad (3)$$

where $P(u_j|q)$ expresses the probability that user u_j answers question q, as estimated by the normalized cosine similarity between question q and the multinomial representation of the whole set of tweets posted by user u_j.

• Conditional entropy $H(g|u_k)$ expresses the information of a single user u_k in group g:

$$H(g|u_k) = p(u_k) \cdot [- \sum_{\substack{u_j \in g \\ u_j \neq u_k}} P(u_j|u_k) \cdot log(P(u_j|u_k))] \quad (4)$$

To achieve our objective of building a collaborative group ensuring the authority of each user u_j and his complementarity with respect to each group member $u_{j'}$, we define the collaboration likelihood indicator $\mathcal{L}_{jj'}$ estimated between users u_j and $u_{j'}$ (this notion is detailed in Section 4.2). Therefore, we propose to estimate $P(u_j|u_{j'})$ as follows:

$$P(u_j|u_{j'}) = \frac{\mathcal{L}_{jk}}{sum_{u_k \in g}\mathcal{L}_{kj'}} \quad (5)$$

4.2 Learning the Pairewise Collaboration Likelihood

The collaboration likelihood $\mathcal{L}_{jj'}$ estimates the potential of collaboration between a pair of users u_j and $u_{j'}$ based on authority and complementarity criteria. We build a predictive model that learns the pairwise collaboration likelihood offline, according to the following statements:

• S1: On Twitter, collaboration between users is noted by the "@" symbol [10, 16].
• S2: Trust and authority enable to improve the effectiveness of the collaboration [26].
• S3: Collaboration is a structured search process in which users might or might not be complementary [36, 37].

More formally, in phase A (Figure 1), we aim to develop a predictive model of the collaboration likelihood using the logistic regression algorithm. In order to train the model, we rely on a set of collaboration pairs $P_{jj'}$ of users u_j and $u_{j'}$ which are modeled according to two elements:
1) The collaboration likelihood $\mathcal{L}_{jj'}$ expressed by a boolean indicator characterizing a collaboration initiated by user u_j and directed to user $u_{j'}$. The likelihood value depends on whether the mentioned user $u_{j'}$ provides feedback ($\mathcal{L}_{jj'} = 1$) or not ($\mathcal{L}_{jj'} = 0$) through an interaction directed at initial user u_j. Indeed, according to statement S1, we hypothesize that the likelihood of a pairwise collaboration can be deduced from the social interactions between the users. Moreover, we assume that the collaboration likelihood is valuable if the mentioned user provides feedback (a reply, retweet and/or mention) to the user who initiated the collaboration.

2) The pairwise collaboration $P_{jj'}$ *of users* u_j *and* $u_{j'}$ is represented by a set of features $\mathcal{F} = \{f_1, \cdots, f_m\}$ denoting social and collaboration abilities in accordance with statements S2 and S3. Therefore, we consider two categories of features estimated at the pair level:

- *Authority-based features* aim to measure the trust and the expertise of each user (Statement S2).
- *Complementarity-based features* aim to measure the extent to which collaborators are complementary in regards to the SNQ&A task (statement S3). We consider three complementarity dimensions: (1) topicality, which addresses different content-based aspects of the question, (2) the types of information provided (video, links, images, etc.), which offers a wide range of pieces of information, and (3) opinion polarity, which provides contrastive subjective information.

It is worth mentioning that the offline trained predictive model is re-injected within the collaborative group building algorithm (Section 4.3) through the group-based information gain metric detailed in Equation 1.

4.3 Building the collaborative group of users

Based on a previous finding highlighting that maximizing the group entropy is equivalent to minimizing the information gain [33], we propose an algorithm (Algorithm 1) in phase B (Figure 1) for recommending a collaborative group. Given an initial group of users, the algorithm discards the least informative user step-by-step with the objective of maximizing the group entropy.

Algorithm 1 Social collaborator recommendation

Input: C ; q
Output: g
▷ Initializing the group of candidate users
$\mathcal{T} = RelevantTweets(C, q)$
$U = Authors(\mathcal{T})$
$t = 0$
$g^t = U$
▷ Learning the collaborative group of answerers
$t^* = \arg\max_{t \in [0,\ldots,|U|-1]} \frac{\partial^2 IG_r(g^t,u)}{\partial u^2}|_{u=u^t}$
  Given $u^t = argmin_{u_{j'} \in g^t} IG_r(g^t, u_{j'})$
  And $g^{t+1} = g^t \setminus u^t$
return g^{t*}

Where:
- C: collection of tweets
- $RelevantTweets(C, q)$: Relevant tweets from collection C considering query q
- $Authors(\mathcal{T})$: authors' of tweets belonging to set $\mathcal{T}$

First, we rely on the compatibility of the question topic and the user's skills, inferred from the user's tweet topics, to identify the initial group of candidate users. Of note, group initialization is only guided by topical relevance to the question and completely neglects the answerers' social relationships with the questioner. Because capturing Twitter users' skills relevant to a topic is a difficult task [31], we rely on the assumption that user relevance may be inferred from the relevance of the his/her tweets. For this purpose, we build the set $\mathcal{T}$ of assumed relevant tweets t_i using a tweet retrieval model (e.g., Berberich et al. [4]). The authors of the identified tweets in $\mathcal{T}$ are used to build the initial set U of candidate users u_j.

Next, we propose an iterative algorithm that decrements the set of candidate users by discarding user u_k, who is the least informative user of the group according to the group-based information gain metric (Equation 1). Therefore, for iteration t, collaborative group g^t is obtained as follows:

$$g^t = g^{t-1} \setminus u^{t-1} \quad (6)$$
$$with\ u^{t-1} = \arg \min_{u_{j'} \in g^{t-1}} IG(g^{t-1}, u_{j'})$$

where g^{t-1} and g^t represent group g at iterations $t-1$ and t, respectively. u^{t-1} expresses the user who has the highest likelihood of being decremented to group g^{t-1}, taking into account the group-based information gain criteria.

In identifying the algorithm's convergence point, we should be able to detect the iteration $t-1$, in which the contribution of user u^{t-1} to group g^{t-1} (expressed by $IG(g^{t-1}, u^{t-1})$ is so high that the group would suffer if it did not include the user. One way to confirm this notion is to analyze the difference in information gain when user u^{t-1} is removed from group g^{t-1}. When this difference is higher than the one obtained for other users who were removed at different timestamps, user u^{t-1} should be deeply engaged in the collaborative process, and the collaborative group g^{t-1} should no longer be decremented. At this level, we believe that the collaboration benefits (workers' skill leveraging) do not outweigh the collaboration costs (cognitive effort, which is positively correlated with the group size, as suggested in [11]). In terms of function analysis, this condition can be assimilated into the iteration t, which maximizes the second partial derivative of the information gain provided by the user who is expected to be removed. Assuming that the information gain is dependent on the group entropy which evolves at each iteration and that the information gain might be biased [33], we used the information gain ratio $IG_r(g^t, u)$ $(\forall u \in g^t)$ at each iteration t which normalizes the metric $IG(g^t, u)$ by the group entropy $H(g^t)$, namely $IG_r(g^t, u) = IG(g^t, u)/H(g^t)$.

The optimization problem might be formulated as follows:

$$t^* = \arg \max_{t \in [0,\ldots,|U|-1]} \frac{\partial^2 IG_r(g^t,u)}{\partial u^2}|_{u=u^t} \quad (7)$$

$$Given\ u^t = argmin_{u_{j'} \in g^t} \quad IG_r(g^t, u_{j'}) \quad (8)$$

$$And\ g^{t+1} = \quad g^t \setminus u^t \quad (9)$$

where the first partial derivative is estimated by the difference between $IG_r(g^t, u)$ and $IG_r(g^{t-1}, u)$ with u satisfying Equation 9 respectively for group g^t and g^{t-1}. The second partial derivative is estimated by the differences between the two associated partial derivatives at the first level.

5. EXPERIMENTAL DESIGN

The objective of our evaluation is to measure the impact of our group recommendation model on the effectiveness of a social network question-answering task. For this aim, we consider a SNQ&A setting in which we attempt to evaluate how much the questioner leverages from the collaborative-based answering.

Accordingly, we estimate both the cohesiveness and relevance of the tweets posted by the members of the recommended group. The research questions that guided our evaluation are the following ones:

RQ1: Do the tweets posted by the collaborative group members recommended by the CRAQ allow the building of an answer? (Section 6.1)

RQ2: Are the recommended group-based answers relevant? (Section 6.2)

RQ3: What is the synergic effect of the CRAQ-based collaborative answering methodology? (Section 6.3)

Our evaluation process relies on the CrowdFlower[2] crowdsourcing platform. Below, we describe our experimental evaluation methodology partially inspired by [21].

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

5.1 Dataset acquisition and processing

We use two publicly available[3] [38] crisis-related Twitter datasets in our experiments: (1) the Sandy hurricane dataset, referring to the most destructive hurricane in the United States and representing a natural disaster crisis. The Twitter stream was monitored from 29^{th} to 31^{st} October 2012 using the keywords *sandy, hurricane* and *storm*, and the stream provided a dataset of 4,853,345 English tweets; (2) the Ebola dataset, referring to the virus epidemic in West Africa and representing a public health crisis. This dataset was collected from July 29^{th} to August 28^{th} 2014 using the keyword *ebola* and includes 4,815,142 English tweets. The dataset was cleaned using an automatic methodology proposed by Imran et al. [18] and exploited in [38]. The objective was to distinguish *informative* tweets that were connected to the crisis or assisted in the understanding of the context from *non-informative* tweets (those not related to the crisis). We established a three-step process: (1) building a training dataset, including 1,800 tweets manually labeled by 10 human judges; (2) refining the classification model using a logistic regression that relies on 12 features. The latter are divided into three classes and are based on the tweet content features (e.g., the number of hashtags), typography features (e.g., the number of punctuation characters) and vocabulary features (e.g., the number of terms belonging to the dictionary extracted from tweets classified as *informative* in the training dataset); and (3) dataset filtering, in which we only retained tweets identified as *informative*. Table 1 presents the descriptive statistics of the resulting datasets.

Table 1: **Descriptive statistics of the Sandy and Ebola Twitter datasets**

Collection	Sandy	Ebola
Tweets	2,119,854	2,872,890
Microbloggers	1,258,473	750,829
Retweets	963,631	1,157,826
Mentions	1,473,498	1,826,059
Reply	63,596	69,773
URLs	596,393	1,309,919
Pictures	107,263	310,581

5.2 Question identification

The first step consists in identifying questions raised on Twitter, with particular attention to identifying questions that are characterized by the intent of obtaining an answer. We apply the question identification methodology proposed in [21] by (1) filtering tweets ending with a question mark; (2) excluding mention tweets and tweets including URLs; (3) filtering tweets with question-oriented hashtags as defined in [20] (e.g., *#help, #askquestion, ...*), as these hashtags indicate that the questioner is likely to be receptive to answers provided by non-network members [21]; and (4) excluding rhetorical questions. In the last step, we ask CrowdFlower workers to manually annotate questions to determine whether they are rhetorical. Each annotation is redundantly performed by three crowd workers, and a final label is retained using CrowdFlower's voting formula. For quality control, we include for each task predefined pairs of question and answer as the gold standard for each task. The annotation task cost is 10 cents per question. Of the 87 and 33 question tweets extracted from the Sandy and Ebola datasets, the identification of rhetorical questions leads to 41 and 22 question tweets, respectively. We outline that only one question

[3]https://figshare.com/collections/expac/3283118

obtained a "reply", which is consistent with previous findings about the lack of answers to tweeted questions [21].

5.3 Evaluation protocol

For evaluation purpose, we adopt a cross-validation methodology in which (a) we learn the predictive model of the pairwise collaboration likelihood (Section 4.2) using one of the datasets and (b) we build the collaborative group (Section 4.3) using the other dataset by estimating the collaboration likelihood metric according to the previously learnt predictive model. The features used are presented in Table 2 and follow the assumptions presented in Section 4.2. Two categories of features are used: the first category aims to measure the authority and the trust of users while the second category measures the complementarity between two collaborators. These two categories are distinguished by the computation point of view. Accordingly, the features in each category are estimated as follows:

- *Authority-based features*: we build a unique metric $X_{jj'}$ for a pair of users u_j and $u_{j'}$ based on the users' intrinsic value (X_j and $X_{j'}$, respectively, for users u_j and $u_{j'}$) of a particular feature, which is based on their importance (e.g., the number of followers), engagement (e.g., the number of tweets), and activity within the question topic (e.g., the number of topically-related tweets). Thus, we propose to combine the users' intrinsic value to measure whether these values are in the same range. In other words, we average the metric values of the two collaborators and divide by their standard deviation to reduce the impact of a wide value range. Moreover, to limit the importance of over-socially active users (such as "News" accounts), we transform the mean-standard deviation ratio using the log function:

$$X_{jj'} = log(\frac{\mu(X_j, X_{j'})}{\sigma(X_j, X_{j'})}) \quad (10)$$

- *Complementarity-based features*: we propose to combine each intrinsic value X_j and $X_{j'}$, respectively, of users u_j and $u_{j'}$ (e.g., the number of tweets with video or tweets with positive opinion) by estimating the absolute difference normalized by the sum of the two collaborators:

$$X_{jj'} = \frac{|X_j - X_{j'}|}{X_j + X_{j'}} \quad (11)$$

We note that the topical complementarity feature between users is estimated differently. We computed the Jansen-Shannon distance between the topical-based representation of users' interests obtained through the LDA algorithm.

5.4 Baselines and comparisons

We evaluate the effectiveness of the CRAQ in comparison to five baselines with respect to the following objectives:

• In order to evaluate the effectiveness of our user-driven and collaborative-oriented approach, we compare our proposed model with the **MMR**, the diversity-based method which ranks tweets by combining relevance and diversity criteria [7]. For fair comparison, we build for each question the top-N tweets retrieved to the question by setting N equal to the number of tweets extracted from the collaborative group members recommended by the *CRAQ*.

• To evaluate the effect of the collaborative group building in the CRAQ, we consider **CRAQ-TR**, the method that only considers the initial group of candidate collaborators so skipping the group entropy maximization process. To ensure a fair comparison with the *CRAQ*, we adopt the same approach detailed in Section 4.3, consisting of inferring users from a set of tweets that are ranked using the topic and temporal-oriented tweet ranking model proposed in [4]. We then build collaborative groups with top-ranked users by ensuring that the size of the group resulting from each question equals the size obtained with the *CRAQ*.

Table 2: **Authority and complementarity-based features modeling collaborators**

	Name	Description
Authority	Importance	Number of followers
		Number of followings
		Number of favorites
	Engagement	Number of tweets
	Activity within the topic	Number of topically-related tweets
		In-degree in the topic
		Out-degree in the topic
Complementarity	Topic	Jansen-Shanon distance between topical-based representation of users' interests obtained through the LDA algorithm
	Multimedia	Number of tweets with video
		Number of tweets with pictures
		Number of tweets with links
		Number of tweets with hashtags
		Number of tweets with only text
	Opinion polarity	Number of tweets with positive opinion
		Number of tweets with neutral opinion
		Number of tweets with negative opinion

• To evaluate the effect of our proposed collaborative group-based answering approach with respect to an individual user-oriented one, we compare our proposed model with **U**, the method that only considers the top ranked user provided by the *CRAQ-TR* [4].

• To assess the contribution of the *CRAQ* algorithm with regard to state-of-the art close approaches, we also run the following baselines that include recent related work:

– **SM**, referring to a structure-based model proposed by [6], in which we use the user group extraction methodology based on URL-sharing behaviors. After the community detection step, we employ the highest cohesion metric value [5] to identify the collaborative group to be recommended. This baseline would enable to evaluate the impact of the social network structure in a group building algorithm.

– **STM**, referring to the Topic Sensitive Page Rank (TSPR) model that relies on both structure and topic to identify important entities with respect to a query [14]. For our experiments, we represent a user through a multinomial distribution of terms of his/her published tweets and build a user network using mention relationships. The TSPR model is then applied to extract a ranking of users. Similar to *CRAQ-TR*, for a given question, the final group includes the same number of users as that in the group recommended by the CRAQ. This baseline would enable to evaluate the impact of collaborative assumptions supporting the CRAQ since this baseline only includes structure and topical assumptions in the group building algorithm.

5.5 Evaluation workflow and Ground truth

Two main tasks are performed by the crowd workers and each task is compensated 35 cents per question. For greater reliability, each task is performed by three crowd workers, and the final label is chosen using the CrowdFlower's voting

Table 3: **Examples of tweet questions and crowd-built answers**

Question	Top ranked tweets of recommended group members	Answer built by the crowd
How do you get infected by this Ebola virus though?? #Twoogle	- http://t.co/D9zc2ZE3DL "@user1: What's this Ebola #Twoogle" - You can get Ebola though Food. By eating infected bats, monkeys, contaminated food.#EbolaFacts - @user2 #Ebola Its shocking how fake news spread fast.. - @user3_ You can get Ebola though Food. By eating infected bats, monkeys, contaminated food.#EbolaFacts	you get ebola by contact with the bodily fluids from an infected person. You can get Ebola though Food. By eating infected bats, monkeys, contaminated food.
Would love to #help to clear up the mess #Sandy made. Any way people can help? Voluntery groups?	- My prayers go out to those people out there that have been affected by the storm. #Sandy - Makes me want to volunteer myself and help the Red Cross and rescue groups.#Sandy - Rescue groups are organized and dispatched to help animals in Sandy's aftermath. You can help by donating. #SandyPets - ASPCA, HSUS, American Humane are among groups on the ground helping animals in Sandy's aftermath. Help them with a donation. #SandyPets #wlf	Rescue groups are organized and dispatched ASPCA, HSUS, American Humane, Donate to @RedCross, @HumaneSociety, @ASPCA.

formula. For quality control, we include for each task predefined pairs of question and answer as the gold standard. Here, we list the crowd evaluation tasks guided by research questions RQ1-RQ3.

Task 1: Answer bulding.

To answer **RQ1**, crowd workers are given the question tweet and the top K posted tweets issued from our CRAQ model and the baselines (with $K = N$ for the MMR - see Section 5.4, and $K \leq 3 * |g|$ for $CRAQ$, $CRAQ - TR$, U, SM, and STM depending on the user's social activity w.r.t. the question topic). They then receive instructions to (a) assess the complementarity and relevance of the recommended group tweets as a whole; rates are included within a range scale 0-3 (0: Not related - 1: Related but not helpful - 2: Related, helpful but redundant - 3: Related, helpful, and complementary), (b) select among the suggested tweets those that aid in formulating an answer, and (c) build an answer using the selected tweet set. Table 3 shows examples of questions, top-ranked tweets posted by the recommended group members and the related crowd-built answers.

Task 2: Relevance assessment.

This task allows the establishment of the ground truth and answering **RQ2-RQ3**. First, for each tweeted question, we identify the recommended group using six settings: (1) the CRAQ and (2) each of the five baseline models. Then, we build interleaved rankings of 20 tweets for each question tweet using (a) the tweet rankings resulting from each model, based on tweets posted by each group member and (b) answers built by crowd workers participating in Task 1 for each setting. Each interleaved ranking is shown to crowd workers who are asked to assess the relevance of both the tweets and the answers built by other crowd-workers (Task 1), using a scale from 0 to 2 (0: "Not relevant" - 1: "Partly relevant" - 2: "Relevant"). Finally, in order to fit with the standard evaluation metrics (precision, recall, F-measure) requiring binary relevance indicators (RQ3), we build the ground truth by gathering for each question all the relevant and partly relevant tweets retrieved by the MMR and authored by users belonging to the collaborative groups recommended by the $CRAQ$ and the remaining baselines.

6. RESULTS AND DISCUSSION

We present here qualitative and quantitative analysis of the CRAQ. We outline that we fixed $N = 100$ (phase A, Figure 1) resulting in groups U with size $|U| \in [25..94]$, since several retrieved tweets might be posted by the same user. Algorithm 1 provided recommended groups $g \subset U$ with size $|g| \in [2..11]$ with only 1 group including more than 5 collaborators.

6.1 Answer building

We start by addressing **RQ1** and test whether (1) the $CRAQ$ is effective in providing useful tweets in terms of relatedness to the question topic and complementarity (Table 4) and (2) those tweets allow building a cohesive answer (Table 5).

In Table 4, we see that tweets of users belonging to the groups recommended by the $CRAQ$ are generally related to the topic of the question and are helpful in answering the question. Indeed, the $CRAQ$ obtains the lowest rate for the "Not related" category, with 12.20% and 7.58% for the Sandy and Ebola datasets, respectively, in comparison with the baseline models, where the proportion of "Not related" tweets varies between 12.20% and 50.41% for the Sandy dataset and 9.09% and 66.70% for the Ebola dataset. Moreover, if we aggregate the results obtained for categories 2 and 3 ("2+3 Related and helpful"), the $CRAQ$ also obtains the highest proportion, with 44.72% and 37.88% for Sandy and Ebola, respectively, with values ranging between 26.02% and 44.15% for Sandy and between 12.12% and 34.85% for Ebola ($p < 0.05$ for all baselines). However, in terms of complementarity, we can see that the $CRAQ$ does not obtain higher results than those obtained by the baselines, with rates equal to 20.33% and 10.61% for both datasets, respectively. This result might be explained by the fact that the collaborative feature-based model was learnt (phase A, Section 4.2) from social interactions between pairwise users who were likely in a local neighborhood context and consequently, similar to each other [27, 21]. To gain a deeper understanding of this observation, we computed the significant complementarity features and related regression estimate values using the logistic regression modeling the collaboration likelihood. As can be seen in Table 6, most of complementarity-based features are significant and all the regression estimate values are negative. The latter observation confirms our previous expectation about the low-level of complementarity used as evidence for learning the collaboration likelihood.

However, by analyzing whether crowd workers are able to formulate an answer based on the users' tweets, Table 5 highlights that this lack of tweet complementarity does not impact on the ability of the recommended group to answer the query. Indeed, the $CRAQ$ achieves one of the highest rates of selected tweets (36.94% for Sandy and 27.42% for

Table 4: **Relatedness and complementarity of the CRAQ results. 0: Not related, 1: Related but not helpful, 2: Related, helpful but redundant, 3: Related, helpful, and complementary, 2+3: Related and helpful tweets**

	Sandy dataset						Ebola dataset					
	MMR[7]	U[4]	$CRAQ$-TR[4]	SM[6]	STM[14]	$CRAQ$	MMR[7]	U[4]	$CRAQ$-TR[4]	SM[6]	STM[14]	$CRAQ$
0	50.41%	43.90%	20.33%	17.07%	12.20%	12.20%	66.70 %	27.72%	19.70%	9.09%	12.12%	7.58%
1	23.58 %	26.82%	38.21%	47.97%	53.66%	43.09%	21.21%	50%	45.45%	59.09%	54.55%	54.55%
2	13.82 %	17.07%	21.14%	17.07%	13.82%	24.39%	9.09 %	13.63%	13.64%	19.70%	13.64%	27.27%
3	12.20 %	12.09%	20.33%	17.89%	20.33%	20.33%	3.03 %	13.63%	21.21%	12.12%	19.70%	10.61%
2+3	26.02 %	29.16%	41.47%	34.96%	44.15%	44.72%	12.12 %	27.26%	34.85%	31.82%	33.34%	37.88%

Table 5: **Answer generation feasibility using the recommended users' tweets.**

	Sandy dataset						Ebola dataset					
	MMR[7]	U[4]	$CRAQ$-TR[4]	SM[6]	STM[14]	$CRAQ$	MMR[7]	U[4]	$CRAQ$-TR[4]	SM[6]	STM[14]	$CRAQ$
Avg percentage of selected tweets	14.12%	37%	24.08%	19.81%	32.79%	36.94%	12.12%	34.85%	24.01%	17.27%	26.52%	27.42%
# built answers	43	29	75	74	67	77	22	11	39	30	37	41

Table 6: **Analysis of the predictive features of collaboration likelihood.** $-: p > 0.05$, $***: p \leq 0.001$

	Feature	Sandy Regres. estimate	Ebola Regres. estimate
Complementarity	Topical	-	-0.23***
	Images	-	-0.19***
	Links	-0.08***	-
	Hashtags	-0.14***	1.58***
	Only text	-0.08***	-0.90***
	Positive opinion	-0.12***	-1.33**
	Neutral opinion	-0.16***	-
	Negative opinion	-0.16***	-

Ebola, with $p < 0.01$ for MMR, $CRAQ$-TR, and SM) and the highest number of built answers (77 and 41 for both datasets, respectively). We outline that the U baseline, although characterized by the highest rate of selected tweets (not significantly different from the $CRAQ$), was not able to select tweets allowing crowd-workers to compose an answer compared to the $CRAQ$ (29 vs. 77, 11 vs. 41, $p < 0.001$ for both Sandy and Ebola datasets respectively). For reminder, each pair of question-users' tweets (respectively 41 and 22 for Sandy and Ebola datasets) has been analyzed by three crowd-workers, leading respectively to 123 and 66 possible built answers for Sandy and Ebola datasets. To summarize, although the CRAQ is not the most effective scenario for selecting complementary tweets, it provides useful information nuggets allowing to build an overall answer to the tweeted question. The key question about the relevance of those answers is addressed in the following section.

6.2 Relevance of built answers

Addressing **RQ2**, Table 7 shows the results of the assessment of the relevance of crowd-built answers based on the tweets provided by the automatically recommended group of users using the $CRAQ$ algorithm and the five baselines. We highlight that χ^2-statistical tests performed on the number of built answer distribution revealed significant differences between the $CRAQ$ and most of the baselines with $p < 0.001$; except for SM baseline for the Ebola dataset where $p > 0.05$.

We can see from Table 7 that compared to the five baselines, the $CRAQ$ enables to build a higher number of answers, among them a higher proportion of "Partly relevant" and "Relevant" answers. For instance, 77.92% (resp. 75.61%) of the $CRAQ$ answers are considered as relevant while the U baseline obtained 68.97% (resp. 72.72%) for the Sandy dataset (resp. Ebola). This statement reinforces our intuition that a single user might have an insufficient knowledge (even if related) to solve a tweeted question. We outline that although the relevance distribution between the MMR and the $CRAQ$ are relatively similar, the number of built answers remains very low for the MMR (half-lower than the $CRAQ$). These results give rise to the benefit of building answers from the user's perspective rather than the tweets regardless of their context.

More particularly, in order to highlight the effect of the CRAQ algorithm, we focus here on the $CRAQ$-TR which is, moreover, the best baseline in terms of the number of built answers and presents an equivalent number of answers as obtained by the $CRAQ$ algorithm. A correlation analysis between the number of built answers and the relevance of those answers reveals that simply recommending authors of top-ranked tweets (the $CRAQ$-TR) enables to select tweets related to the question, but mostly not relevant. Indeed, for the Sandy dataset (resp. Ebola), the correlation analysis indicates a non-significant coefficient value of 0.31 (resp. 0.20) for the $CRAQ$-TR and, unlikely, a significant coefficient of 0.89 with $p < 0.001$ (resp. 0.84, $p < 0.001$) for the $CRAQ$ algorithm. In the same mind, we notice that the $CRAQ$ is characterized by a higher number of built answers which have been assessed as "(Partly) relevant" (2+3) than the $CRAQ$-TR (resp. 60 vs. 52 for Sandy and 31 vs. 24 for Ebola). Intuitively, this difference could be explained by the crowd workers' assessments of the complementarity and relatedness of group tweets, as the $CRAQ$ obtained a lower rate of non-related recommended groups than the $CRAQ$-TR (e.g., 12.20% vs. 20.33% for the Sandy dataset). These observations suggest that building a group by gathering individual users identified as relevant through their skills (tweet topical similarity with the question) is not always appropriate. This finding emphasizes the benefit of recommending groups in which users taken as a whole contribute to the group entropy based on their social activity in addition to their skills.

6.3 Synergic effect of CRAQ-based collaborative answering methodology

Turning to **RQ3**, we aim to measure the synergic effect of simulated collaboration within the recommended groups of users with respect to the search effectiveness based on the tweets published by those group members (the $CRAQ$ algorithm and the five baselines). Effectiveness is measured using traditional precision, recall and F-measure metrics. Statistical significance of observed differences between the performance of compared runs is tested using a two-tailed

Table 7: **Analysis of the relevance of the crowd-built answers. ba: Number of built answers, 1: Not relevant, 2: Partly relevant, 3: Relevant, 2+3: (Partly) Relevant**

		MMR[7]	*U*[4]	*CRAQ-TR*[4]	*SM*[6]	*STM*[14]	*CRAQ*
Sandy	ba	43	29	75	74	67	77
	1	11: 25.58%	9: 31.03%	23: 30.67%	24: 32.43%	21: 31.34%	17: 22.08%
	2	20: 46.51%	14: 48.28%	33: 44.00%	34: 45.95%	24: 35.82%	39: 50.65%
	3	12: 27.91%	6: 20.69%	19: 25.33%	16: 21.62%	22: 32.84%	21: 27.27%
	2+3	32: 74.42%	20: 68.97%	52: 69.33%	50: 67.57%	46: 68.66%	60: 77.92%
Ebola	ba	22	11	39	30	37	41
	1	4: 21.05%	3: 27.27%	15: 38.46%	8: 26.67%	15: 40.54%	10: 24.39%
	2	6: 31.58%	4: 36.36%	18: 46.15%	15: 50%	16: 43.24%	22: 53.66%
	3	9: 47.37%	4: 36.36%	6: 15.38%	7: 23.33%	6: 16.22%	9: 21.95%
	2+3	15: 78.95%	8: 72.72%	24: 1.53%	22: 73.33%	22:59.46%	31: 75.61%

Table 8: **Effectiveness analysis at the tweet level for recommended collaborative groups. %Chg: *CRAQ* improvement. Significance t-test:** $*: 0.01 < \alpha \leq 0.05$, $**: 0.001 < \alpha \leq 0.01$, $***: \alpha \leq 0.001$

	MMR [7]		*U* [4]		*CRAQ-TR* [4]		*SM* [6]		*STM* [14]		*CRAQ*
	Value	%Chg	Value	%Chg	Value	%Chg	Value	%Chg	Value	%Chg	**Value**
Sandy dataset											
Precision	0.24	+92.93**	0.46	+2.32	0.33	+42.01*	0.21	+124.71***	0.49	-4.1	0.47
Recall	0.09	+95.19*	0.1	+81.63*	0.15	+16.18	0.09	+105.96*	0.1	+80.09*	0.18
F-measure	0.12	+78.22*	0.15	+41.79	0.19	+10.59	0.12	+84.42*	0.15	+40.48	0.21
Ebola dataset											
Precision	0.22	+153.65***	0.64	-12.12	0.5	+12.22	0.3	+89.59**	0.45	+24.5	0.57
Recall	0.07	+155.59***	0.11	+69.96*	0.22	-18.08	0.12	+46.80	0.06	+216.56***	0.18
F-measure	0.09	+164.17***	0.21	+17.46	0.28	-11.64	0.17	+50.07	0.1	+159.07***	0.25

paired t-test. We can see from Table 8 that according to the recall, the *CRAQ* displays significantly better results for the Sandy dataset, between +80.09% and +105.95%, compared with the *MMR*, *U*, *SM*, and *STM* baseline models and for the Ebola dataset, between +69.96% and +216.56%, compared with the *MMR*, *U*, and *STM* baseline models. Specifically, we can highlight the following statements:

• The *CRAQ* significantly overpasses the *MMR* baseline with improvements ranging from 78.22% to 164.17% over both datasets. This sustains the statement observed in the qualitative analysis and reinforces our intuition to also consider topical diversity from the user context perspective in addition to the content perspective.

• The comparison with the *U* baseline, referring to the recommendation of a single collaborator, highlights that the *CRAQ* is able to obtain similar precision values while significantly increasing the recall. These results are consistent with previous work [35, 37], highlighting the synergic effect of a collaborative group in which the collaborative effort of several users is able to provide a larger amount of relevant information without hindering the precision.

• The *CRAQ* generally outperforms the *CRAQ-TR*, with a significant improvement of the precision metric for the Sandy dataset. Moreover, we highlight that non-significant performance changes are observed for the Ebola dataset. Combining these results with those of the qualitative analysis of the relevance of crowd-built answers (Section 6.2) indicate that although the levels of recall of both the *CRAQ* and *CRAQ-TR* are similar, the *CRAQ* algorithm allows building a higher ratio of relevant answers. Considering the fact that the *CRAQ-TR* only relies on topical relevance to build the group of answerers, this result could be explained by the key feature of the *CRAQ* algorithm which consists in maximizing group entropy based on users' collaboration likelihood.

• The *CRAQ* overpasses the *SM* with significant improvements from +84.42% to +124.71% over the three evaluation metrics and both datasets. We recall that while the *SM* baseline model relies on strong weak ties underlying users' local social network to select the potential answerers, the *CRAQ* algorithm considers topical matching between the question and the user's tweets without considering the strength of their social ties. Hence, this result corroborates our belief that a questioner might benefit from collaboration with relevant strangers.

• The *CRAQ* algorithm significantly outperforms the *STM* for both datasets with respect to the recall and F-measure metrics. This baseline achieves very small recall values (0.1 and 0.06 for the Sandy and Ebola datasets, respectively), suggesting that topically relevant tweets issued from the most socially authoritative are not obviously relevant to answer the tweeted question. This finding reinforces our driving idea around collaborative-based answering and is consistent with previous findings which claim that collaborative groups might be built based on user's skills regarding the task at hand and user-user affinity likelihood in order to ensure the synergic effect of collaboration [34, 35].

7. CONCLUSION AND FUTURE WORK

We have presented a novel method for answering questions on social networks. The proposed method fundamentally relies on the assumption that a relevant answer is an aggregation of topically relevant posts provided by a collaborative-based approach built on a group of potential answerers. Therefore, the key question is how to build the optimal group of candidate answerers. The CRAQ is an attempt to answer this question. It iteratively builds candidate collaborative groups of users that are subject to key properties: the information gain that an optimal user supplies to his/her group, complementarity and topical relevance of the related tweets, and trust and authority of the group members. Our experimental evaluation on two Twitter datasets demonstrated the effectiveness of our proposed algorithm. Although we focused on questions posted on Twitter, our method is applicable to other social platforms, such as Face-

book and community Q&A sites. One limitation is that the predictive model of collaboration likelihood relies on basic assumptions of collaboration (e.g., mention). A deeper analysis of collaboration behavior on social networks would help to identify collaboration patterns, and then better estimate the collaboration likelihood. The relevance ratings of both tweets posted by the CRAQ-based recommended users and the related manually built answers are very promising, and we plan to extend this work through the automatic summarization of candidate answers.

8. ACKNOWLEDGMENTS

This research was supported by the French CNRS PEPS research program under grant agreement EXPAC (CNRS/PEPS 2014-2015).

9. REFERENCES

[1] I. Abraham, O. Alonso, V. Kandylas, R. Patel, S. Shelford, and A. Slivkins. Using worker quality scores to improve stopping rules. In *SIGIR*. ACM, 2016.

[2] L. E. Agustín-Blas, S. Salcedo-Sanz, E. G. Ortiz-García, A. Portilla-Figueras, A. M. Pérez-Bellido, and S. Jiménez-Fernández. Team formation based on group technology: A hybrid grouping genetic algorithm approach. *Computers & Operations Research*, 38(2):484 – 495, 2011.

[3] G. Barbara. *Collaborating: finding common ground for multiparty problems*. Jossey-Bass, 1989.

[4] K. Berberich and S. Bedathur. Temporal Diversification of Search Results. In *SIGIR #TAIA workshop*. ACM, 2013.

[5] R. D. Bock and S. Z. Husain. An adaptation of holzinger's b-coefficients for the analysis of sociometric data. *Sociometry*, 13(2):pp. 146–153, 1950.

[6] C. Cao, J. Caverlee, K. Lee, H. Ge, and J. Chung. Organic or organized?: Exploring url sharing behavior. In *CIKM*, pages 513–522. ACM, 2015.

[7] J. Carbonell and J. Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *SIGIR*, pages 335–336. ACM, 1998.

[8] D. Castilho, P. Melo, D. Quercia, and F. Benevenuto. Working with friends: Unveiling working affinity features from facebook data. In *ICWSM*, 2014.

[9] S. Chang and A. Pal. Routing questions for collaborative answering in community question answering. In *ASONAM*, pages 494–501. ACM, 2013.

[10] K. Ehrlich and N. S. Shami. Microblogging Inside and Outside the Workplace. In *ICWSM*, 2010.

[11] R. Fidel, A. M. Pejtersen, B. Cleal, and H. Bruce. A Multidimensional Approach to the Study of Human-information Interaction: A Case Study of Collaborative Information Retrieval. *JASIST*, 55(11):939–953, 2004.

[12] J. Foster. Collaborative information seeking and retrieval. *Annual Review of Information Science and Technology*, 40(1), 2006.

[13] R. González-Ibáñez, C. Shah, and R. W. White. Capturing collabportunities: A method to evaluate collaboration opportunities in information search using pseudocollaboration. *JASIST*, 66(9):1897–1912, 2015.

[14] T. H. Haveliwala. Topic-sensitive pagerank. In *WWW*, pages 517–526, 2002.

[15] B. Hecht, J. Teevan, M. R. Morris, and D. J. Liebling. SearchBuddies: Bringing Search Engines into the Conversation. In *ICWSM*, 2012.

[16] C. Honey and S. Herring. Beyond Microblogging: Conversation and Collaboration via Twitter. In *HICSS*, pages 1–10, 2009.

[17] D. Horowitz and S. D. Kamvar. The Anatomy of a Large-scale Social Search Engine. In *WWW*, pages 431–440, 2010.

[18] M. Imran, S. Elbassuoni, C. Castillo, F. Diaz, and P. Meier. Extracting information nuggets from disaster- related messages in social media. In *ISCRAM*, pages 791–801, 2013.

[19] R. Jeffrey M., S. Emma S., M. Jorge Nathan, M.-H. Andres, and M. R. Morris. Is anyone out there? unpacking q&a hashtags on twitter. In *CHI*, 2014.

[20] R. Jeffrey M and M. R. Morris. Estimating the social costs of friendsourcing. In *CHI*, 2014.

[21] J.-W. Jeong, M. R. Morris, J. Teevan, and D. J. Liebling. A Crowd-Powered Socially Embedded Search Engine. In *ICWSM*, 2013.

[22] B. Li and I. King. Routing questions to appropriate answerers in community question answering services. In *CIKM*. ACM, 2010.

[23] H. Li, B. Zhao, and A. Fuxman. The wisdom of minority: Discovering and targeting the right group of workers for crowdsourcing. In *WWW*, pages 165–176, 2014.

[24] Y. Liu, J. Bian, and E. Agichtein. Predicting information seeker satisfaction in community question answering. In *SIGIR*, pages 483–490. ACM, 2008.

[25] J. Mahmud, X. Zhou, Michelle, N. Megiddo, J. Nichols, and C. Drews. Recommending targeted strangers from whom to solicit information on social media. In *IUI*, pages 37–48.

[26] K. McNally, M. P. O'Mahony, and B. Smyth. A model of collaboration-based reputation for the social web. In *ICWSM*, 2013.

[27] M. McPherson, L. S. Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1):415–444, 2001.

[28] M. R. Morris. Collaborative Search Revisited. In *CSCW*, pages 1181–1192, 2013.

[29] M. R. Morris, J. Teevan, and K. Panovich. What do people ask their social networks, and why?: A survey study of status messages. In *CHI*, pages 1739–1748, 2010.

[30] A. Oeldorf-Hirsh, M. R. Morris, J. Teevan, and G. Darren. To search or to ask: The routing of information needs between traditional search engines and social networks. In *ICWSM*, 2014.

[31] A. Pal and S. Counts. Identifying topical authorities in microblogs. In *WSDM*, pages 45–54, 2011.

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

[33] J. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.

[34] H. Rahman, S. B. Roy, S. Thirumuruganathan, S. Amer-Yahia, and G. Das. "The Whole Is Greater Than the Sum of Its Parts": Optimization in Collaborative Crowdsourcing. *CoRR*, abs/1502.05106, 2015.

[35] C. Shah and R. González-Ibáñez. Evaluating the synergic effect of collaboration in information seeking. In *SIGIR*, pages 913–922, 2011.

[36] D. H. Sonnenwald. Communication roles that support collaboration during the design process. *Design Studies*, 17(3):277–301, July 1996.

[37] L. Soulier, C. Shah, and L. Tamine. User-driven system-mediated collaborative information retrieval. In *SIGIR*, pages 485–494. ACM, 2014.

[38] L. Tamine, L. Soulier, L. B. Jabeur, F. Amblard, C. Hanachi, G. Hubert, and C. Roth. Social media-based collaborative information access: Analysis of online crisis-related twitter conversations. In *HT*, pages 159–168. ACM, 2016.

[39] J. C. Tang, M. Cebrian, N. A. Giacobe, H.-W. Kim, T. Kim, and D. B. Wickert. Reflecting on the DARPA Red Balloon Challenge. *Commun. ACM*, 54(4):78–85, 2011.

[40] J. Teevan, D. Ramage, and M.-R. Morris. Twitter search: A comparison of microblog search and web search. In *WSDM*, pages 35–44, 2011.