Retrieving Relevant Conversations for Q&A on Twitter

Jose Miguel Herrera
PRISMA Research Group
Department of Computer Science
University of Chile, Chile
jherrera@dcc.uchile.cl

Denis Parra
Department of Computer Science
Pontificia Universidad Católica de Chile
dparra@ing.puc.cl

Barbara Poblete
PRISMA Research Group
Department of Computer Science
University of Chile
bpoblete@dcc.uchile.cl

ABSTRACT

Community Question and Answering (Q&A) sites provide special features for asking questions and receiving answers from users on the Web. Nevertheless, Web users do not restrict themselves to posting their questions exclusively in these platforms. With the massification of on-line social networks (OSN) such as Twitter, users are increasingly sharing their information needs on these websites. Their motivation for doing so is to obtain a timely and reliable answer from their personal community of trusted contacts. Therefore, daily on Twitter, there are hundreds of thousands of questions being shared among users from all over the world. Many of these questions go unanswered, but also an important number receive relevant and complete replies from the network. The problem is that due to the volatile nature of the streaming data in OSN and the high arrival rates of messages, valuable knowledge shared in this Q&A interaction lives very shortly in time. This produces high redundancy and similarity in questions which occurs consistently over time. Following this motivation we study Q&A conversations on Twitter, with the goal of finding the most relevant conversations posted in the past that answer new information needs posted by users. To achieve this we create a collection of Q&A conversation threads and analyze their relevance for a query, based on their contents and relevance feedback from users. In this article, we present our work in progress which includes a methodology for retrieving and ranking Q&A conversation threads for a given query. Our preliminary findings show that we are able to use historical conversation on Twitter to answer new queries posted by users. We observe that in general the asker's feedback is a good indicator of thread relevance. Also, not all of the feedback features provided by Twitter are equally useful for ranking Q&A thread relevance. Our current work focuses on determining empirically the best ranking strategy for the recommendation of relevant threads for a new user question. In the future we seek to create an automatic Q&A knowledge base that is updated in real-time that allows for preserving and searching human understanding.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Relevance feedback; H.1.2 [**User/Machine Systems**]: Human factors

General Terms

Experimentation, Human Factors, Algorithms.

Keywords

Twitter, Recommendation, Ranking, Q&A, Threads.

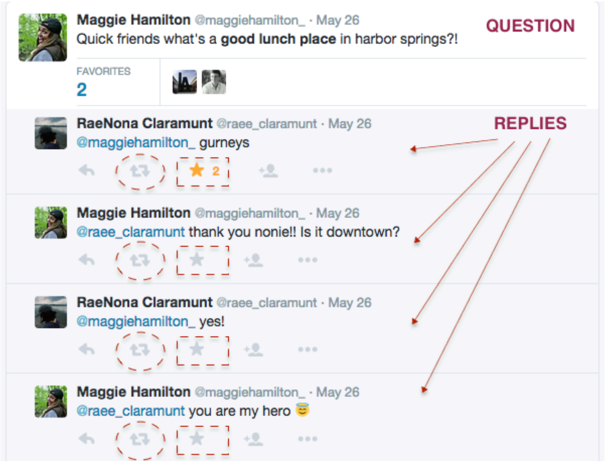
1. INTRODUCTION

Question and Answering (Q&A) websites allow users to ask questions and receive answers from a diverse group of people. These kinds of sites accumulate knowledge providing a valuable resource of information that cannot be easily obtained using Web search engines. Popular Q&A sites are *Yahoo! Answers*[1] and *Stack Overflow*[2], which are specially designed to generate Q&A interaction among users. One property of these kinds of sites is the ability to choose the *best answer* for a particular question through a community-wise voting system. In general, the *best answer* selection is based on the amount of positive votes for a user's answer. This feature is one of the most important characteristics in these kinds of sites, because these highly voted answers will solve similar future questions. With this social mechanism that allows for collecting good-quality questions and answers, Q&A sites motivate people to come back for obtaining almost immediate answers to their information needs.

Although Q&A sites are popular, users also ask a significant volume of questions online in other non-specialized but more popular networking platforms, such as Facebook or Twitter. This behavior might seem counterintuitive, especially on Twitter, due to the volatile nature of its information stream and the lack of special incentives for motivating users' answers available in Q&A sites – such as badges and enhanced rights for active users. This seemingly suboptimal behavior might be explained by the observations of Morris et al. [7], who showed that the main motivation that users have for asking questions online on Q&A platforms is to receive quick and trustworthy answers - something that can be potentially achieved in a massive microblogging site like Twitter. Considering this background, we conducted a preliminary analysis which indicated that around 10% of the Twitter stream corresponds to Q&A messages (similar measures were obtained by [4, 6]). Moreover, a rough inspection at Q&A conversation threads on Twitter yields high redundancy of questions over time, meaning that there is a high chance of finding answers to newly asked questions. This trend of Q&A usage in Twitter indicates a increasing potential for fulfilling current users' information needs based on similar questions already answered in the past. Previous work shows research

[1] http://answers.yahoo.com: General-purpose Q&A.

[2] http://www.stackoverflow.com: Software-development Q&A.



Maggie Hamilton @maggiehamilton_ · May 26
Quick friends what's a **good lunch place** in harbor springs?!
FAVORITES
2

RaeNona Claramunt @raee_claramunt · May 26
@maggiehamilton_ gurneys
★ 2

Maggie Hamilton @maggiehamilton_ · May 26
@raee_claramunt thank you nonie!! Is it downtown?

RaeNona Claramunt @raee_claramunt · May 26
@maggiehamilton_ yes!

Maggie Hamilton @maggiehamilton_ · May 26
@raee_claramunt you are my hero 😇

Figure 1: A conversation thread on *Twitter* formed by one question and four answers. Tweets can be *Re-tweeted* (dotted circle) and *Favorited* (dotted square).

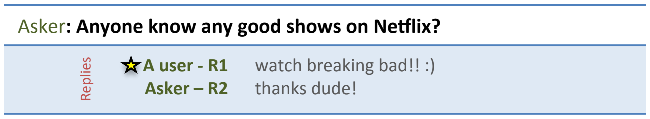
on conversations in Twitter, such as [1, 2, 5], but none of them focuses on building a knowledge repository of Q&A in Twitter to answer a question's query. Other researchers have tried to automatically reply to unanswered questions by matching similar questions already answered in the past [8, 10], but they employed the *Yahoo! Answers* platform, and the challenge in Twitter is more complex due to the lack of explicit mechanisms to tell good from bad answers as in Q&A sites. For these reasons, we address the task of creating a method for retrieving the most relevant historical conversation threads which can answer a given query q, by leveraging Twitter as Q&A repository.

We study how the combination of questions, their replies and Twitter social features (retweets, replies, favorites, etc.) can be useful for determining whether a conversation thread triggered by a question is relevant in terms of information quality, to the particular conversation topic. In particular, we aim to investigate the following research questions:

- **RQ1:** how can we determine whether a conversation thread was resolved (answered) on Twitter? In other words, which factors or features are most relevant to determine that? and,
- **RQ2:** can we recommend relevant conversation threads made in the past to answer a new question? Can we build a ranking model with the most relevant features of **RQ1**?

We define the relevance of a conversation thread in terms of its likelihood of providing a complete answer and resolving the information need. Then, to evaluate the importance of each conversation we employed a relevance measure which is based on the feedback provided by the user asking the question on Twitter. Our preliminary findings show that the feedback of the asker plays an important role to evaluate whether a conversation thread had good quality, i.e., whether it was resolved or not. Nevertheless, the noisy nature of tweets makes them complex to analyze, making our problem difficult.

The main contributions of this paper are: **(1)** proposing a methodology for obtaining a set of ranked historical conversation threads that answer a given question, and **(2)** identifying the main characteristics that influence the quality of a conversation thread. To the best of our knowledge, the method proposed in this article is the first attempt to rank conversation threads based on feedback in Twitter. The applications of this work can be useful for any social network with interaction among users to enhance the results on search. Also, we can use this approach to build a question and answer repository website based on Twitter.



Asker: **Anyone know any good shows on Netflix?**

☆**A user - R1** watch breaking bad!! :)
Asker – R2 thanks dude!

Figure 2: An example of a thread with two replies (R1 and R2) where an *Asker* asks a question, a *User* replies, and finally the *Asker* replies back. The star means the *Asker* has marked that reply as Favorite.

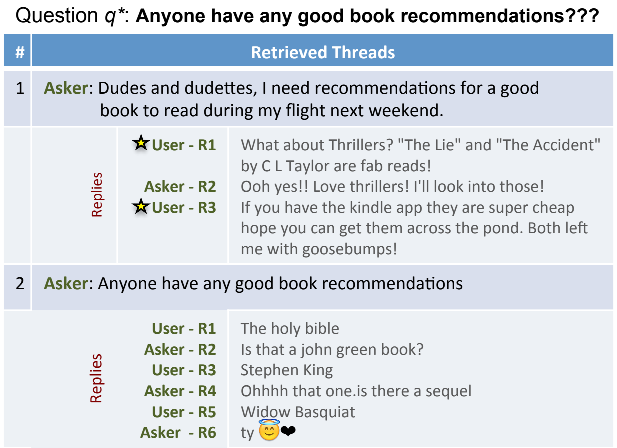
The rest of the paper is structured as follows: Section 2 describes our methodology to identify and rank Q&A threads based on questions asked on Twitter, Section 3 provides details of our preliminary experiments, such as the dataset and uses cases where our model works appropriately, and Section 4 provides a brief summary of our initial expectations and future work.

2. RANKING Q&A THREADS

In this section we present our preliminary methodology for retrieving and ranking relevant Q&A conversation threads for a previously unknown query on Twitter. We define a Q&A conversation thread (hereinafter *threads*), as a conversation on Twitter in which the initial tweet is phrased as a question. We define the relevance of a thread in terms of *how effective the complete conversation is at answering the information need posed in the initial tweet of the conversation*. See Figure 1 for an example of a Q&A conversation thread.

In order to identify features that characterize whether a conversation thread has already resolved an information need, we manually inspected several hundreds of conversation threads. This analysis brings us to consider that replies in conversation threads are important at the moment of establishing the relevance of the conversation. In particular, given Twitter's relevance feedback options, tweets in a conversation thread can be marked by users as *Favorite* and/or *Retweeted*, where the first indicates a special preference and the second indicates that the content has been re-posted by a user. In particular, we observe that the feedback provided by the user who posted the question which initiates a thread, called the *asker*, plays an important role indicating the relevance of the thread. Our intuition is that since the asker is very interested in obtaining a good answer to his/her query, a frequent use of Twitter's relevance feedback features will indicate a higher satisfaction. Figure 2 shows a simple instance where the asker gives feedback in a thread. In this case, with an option to determine the level of satisfaction of the asker with a thread, we can evaluate the second reply "thanks dude!" of the *asker* using Sentiment Analysis (using NLTK tools[3] we obtain the reply has a positive polarity of $67,13\%$). This follows a similar approach by Pelleg et al. [8] for Yahoo! Answers. In the example, if the asker additionally marks the first reply as a *Favorite* (followed by a positive answer) this provides a stronger

[3] Natural Language Toolkit, http://text-processing.com



Question q*: **Anyone have any good book recommendations???**

#	Retrieved Threads		
1	**Asker**: Dudes and dudettes, I need recommendations for a good book to read during my flight next weekend.		
	Replies	★**User - R1**	What about Thrillers? "The Lie" and "The Accident" by C L Taylor are fab reads!
		Asker - R2	Ooh yes!! Love thrillers! I'll look into those!
		★**User - R3**	If you have the kindle app they are super cheap hope you can get them across the pond. Both left me with goosebumps!
2	**Asker**: Anyone have any good book recommendations		
	Replies	**User - R1**	The holy bible
		Asker - R2	Is that a john green book?
		User - R3	Stephen King
		Asker - R4	Ohhhh that one.is there a sequel
		User - R5	Widow Basquiat
		Asker - R6	ty 😇❤

Figure 3: Case 1. Given a question q^*, we show the top-2 relevant threads. The stars mean that the tweet was marked as Favorite by the Asker.

indication of satisfaction with the reply. We call this type of behavior *positive reinforcement feedback* (PRF), in which the asker indicates its approval for replies to his/her question. In our initial inspection of our dataset we have identified at least 5 other similar types of PRF which are recurrent over time in Q&A threads.[4]

More formally, given a new question q^* we retrieve a set of its top-k similar conversation threads. We do so initially by retrieving threads with the highest cosine similarity of their initial tweet q^*. We denote this set of similar threads as $T = \{th_1, \ldots, th_k\}$. Each thread is given by $th_j =< q_j, R_j >$, in which q_j is the initial tweet or question of th_j and R_j is the set of replies received for q_j. Then, for each thread th_j we compute its absolute relevance $rel(th_j)$ that indicates the level of satisfaction of the asker of q_j with the overall replies received in th_j. Initially we estimate $rel(th_j)$ as:

$$rel(th_j) = count\ of\ positive\ reinforcement\ instances\ in\ th_j$$

Using the value of $rel(th_j)$ we re-order the set T obtained for q^*. Just we take the top-k elements with highest $rel(th_j)$. We do not use a threshold value because each thread presents a different levels of feedback. Hence, we can not define a fixed value.

3. PRELIMINARY EXPERIMENTS

In this section we present the dataset used in our experiments, preliminary results and some findings.

Dataset. In order to show evidence of the usefulness of our proposal, we collected a dataset of tweets in English language. This preliminary Twitter dataset contains 721 questions (q^*) and 152, 721 conversation threads (th_j). We created this collection from the public Twitter API. Since our goal is to have sets of similar questions in the dataset and question threads are very sparse in the public stream, we conducted a focalized crawl for threads. This is, we retrieved questions and similar question threads using the following iterative process: 1) we search in Twitter for a list of common words used in questions, 2) filter all of the results (tweets) that corresponded to a question q^*, and 3) for each question q^*, perform an additional search to retrieve similar questions-threads th_j. The full process (1)-(3) was conducted between March 31, 2015 and April 27, 2015.

The first and second steps were carried out through the **Streaming API**[5] using a traditional rule-based approach [4, 7, 11]. Also, we have defined certain rules of questions that we need, because not just any question is useful in our task. We collect questions that require answers (information needs or factual knowledge), questions that are not affected by time, recommendation questions, suggestion requests, questions in English, and opinion questions. For instance, we keep in our dataset questions such as: “does anyone know cheap places to stay in London?”, “does anybody can recommend me good restaurants in Santa Monica?”. On the other hand, we discard questions such as: “anyone got an iPhone 5 for sale?”, “Anyone know what time the mayweather fight starts??”. The first is not a factual knowledge question and the second is affected by time (we have proposed to include these kinds of questions in future work).

Ranking conversation threads. The third step is to build the set of similar past threads th_j of q^*. Since Twitter API restricts obtaining the complete thread, we must first retrieve similar tweets and then the replies, if they exist. We have retrieved all the q_j that are similar to q^* from the **Search API**[6] (the retrieval is based on the main keywords of q^*). Then, we retrieved the replies R_j of q_j to build the thread structure. We have adapted a development made by Adrian Crepaz[7] that can get replies through the Twitter mobile webpage. Finally, we calculate the relevance $rel(th_j)$ for each thread based on the PRF.

3.1 Q&A Ranking Examples

By both automatic analysis and manual inspection of our dataset, we identified common patterns of Q&A conversation threads. In this subsection we describe three of the most recurrent examples and how our ranking methodology works in each case.

Case 1. Figure 3 shows the top 2 similar threads retrieved by our approach sorted by relevance (high to low), given the initial question q^*: “Anyone have any good book recommendations???” (it was taken literally). The first thread contains three replies and two of them were marked as Favorites by the asker (see the starts). Notice that the first reply R1 (of the first thread) was marked as Favorite by the asker and followed by the asker (R2) with positive sentiment. The sentiment analysis of R2 gave us 76% of probability that the text presents positive polarity. That means that the thread presents PRF. The reply R3 of the first thread was marked as favorite by the asker but it is not followed by any tweet of the asker. On the other hand, the second thread just presents a positive sentiment in the reply R6 (“ty” means “thank you”). Although the asker uses feedback elements (positive expression in the reply R6), the second thread does not present the structure to be PRF. Hence, the relevance is lower than the first thread.

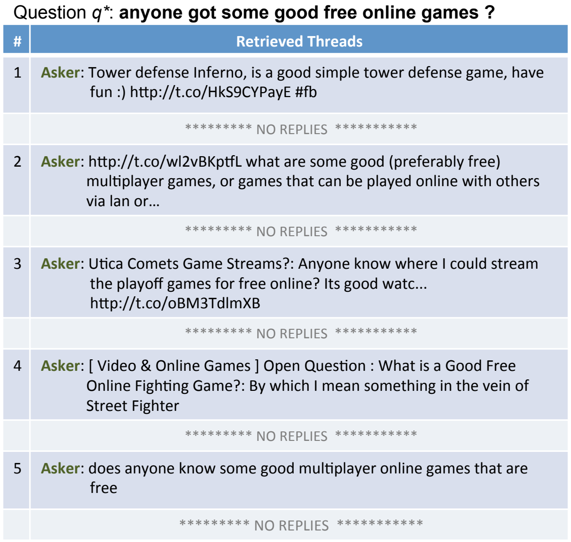
Case 2. Figure 4 presents another recurrent case. Given an initial question q^*, the threads retrieved are just the initial tweet q_i, without replies. But if we observe, the retrieved tweets still can answer the question q^*. When this occurs, we sort the tweets depending

[4] We do not enter in details at this moment of all of the types of PRF, given that we are presenting our preliminary findings in brief format.

[5] The streaming API captures 1% of the Twitter volume in real time.

[6] The search API retrieves tweets posted within a week of the time the query was issued.

[7] http://adriancrepaz.com/twitter_conversations_api



Question *q**: **anyone got some good free online games ?**

#	Retrieved Threads
1	**Asker**: Tower defense Inferno, is a good simple tower defense game, have fun :) http://t.co/HkS9CYPayE #fb
	********* NO REPLIES ***********
2	**Asker**: http://t.co/wl2vBKptfL what are some good (preferably free) multiplayer games, or games that can be played online with others via lan or...
	********* NO REPLIES ***********
3	**Asker**: Utica Comets Game Streams?: Anyone know where I could stream the playoff games for free online? Its good watc... http://t.co/oBM3TdlmXB
	********* NO REPLIES ***********
4	**Asker**: [Video & Online Games] Open Question : What is a Good Free Online Fighting Game?: By which I mean something in the vein of Street Fighter
	********* NO REPLIES ***********
5	**Asker**: does anyone know some good multiplayer online games that are free
	********* NO REPLIES ***********

Figure 4: Case 2. The top-5 threads retrieved are just tweets (without replies), but we can sort them by URLs.

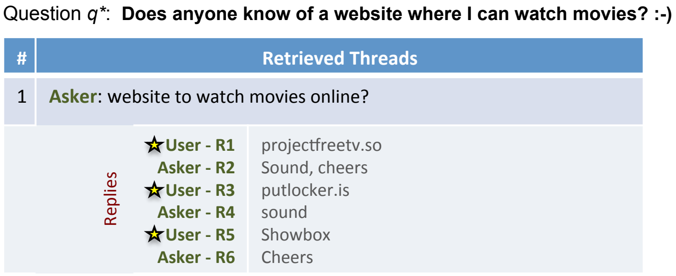
on whether they contain URLs. Chen et al. [3] shows that the tweet relevance is high when it contains a URL. In our dataset, the amount of threads without replies are 69.9%. We notice that after the third tweet the tweets do not clearly reply to the initial question $q*$.

Case 3. The conversation threads could have high relevance if they had more instances of PRF within the same thread. Figure 5 shows this case, where the thread presents PRF twice in one thread. The reply R1 has been marked as Favorite by the Asker. The reply R2 was made by the Asker with 60% of positive sentiment. Hence, the replies R1-R2 present PRF. The replies R3-R4 also present PRF (R3 was marked as Favorite by the Asker and reply R4 returned 54% of positive sentiment). Although reply R5 has a Favorite made by the Asker, the reply R6 has a neutral sentiment. Therefore, they do not present PRF.

4. CONCLUSIONS AND FUTURE WORK

The cases presented in this chapter provide evidence of how our method is used for retrieving and ranking historical conversations threads in order to answer recent questions. This is preliminary work and there is much left to do in the future, such as, validation based on human judgement. The main goals of the evaluation are: (1) whether our positive reinforcement instances are accurate to correctly classify relevant threads (**RQ1**), and (2) whether our ranking approach supports recommendation of relevant threads (**RQ2**).

Such results can lead us to to determining the best model and highlight which Twitter features are more relevant to our task. If we detect that there are several features that can influence in determining the relevance of a thread, we propose use machine learning techniques to automatically construct the ranking model based on the aforementioned features.



Question *q**: **Does anyone know of a website where I can watch movies? :-)**

#	Retrieved Threads	
1	**Asker**: website to watch movies online?	
Replies	★**User - R1**	projectfreetv.so
	Asker - R2	Sound, cheers
	★**User - R3**	putlocker.is
	Asker - R4	sound
	★**User - R5**	Showbox
	Asker - R6	Cheers

Figure 5: Case 3. Two types of *positive reinforcement feedback* (PRF) in one thread: R1 was marked as *Favorite* by asker and R2 has positive sentiment. The same with replies R3-R4. The reply R5 has a favorite but R6 presents a neutral sentiment.

5. REFERENCES

[1] D. Boyd, S. Golder, and G. Lotan. Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter. In *System Sciences (HICSS), 2010 43rd Hawaii International Conference on*, 2010.

[2] J. Chen, R. Nairn, and E. H.-h. Chi. Speak little and well: recommending conversations in online social streams. *CHI*, 2011.

[3] J. Chen, R. Nairn, L. Nelson, M. Bernstein, and E. Chi. *Short and tweet: experiments on recommending content from information streams*. ACM, 2010.

[4] M. Efron and M. Winget. Questions are content: a taxonomy of questions in a microblogging environment. In *ASIS&T '10: Proceedings of the 73rd ASIS&T Annual Meeting on Navigating Streams in an Information Ecosystem*, 2010.

[5] C. Honey and S. C. Herring. Beyond Microblogging: Conversation and Collaboration via Twitter. In *System Sciences, 2009. HICSS '09. 42nd Hawaii International Conference on*, 2009.

[6] B. Li, X. Si, M. R. Lyu, I. King, and E. Y. Chang. Question identification on twitter. In *CIKM '11: Proceedings of the 20th ACM international conference on Information and knowledge management*, 2011.

[7] 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 *CHI '10: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2010.

[8] D. Pelleg, O. Rokhlenko, M. Shovman, I. Szpektor, and E. Agichtein. The Crowd is Not Enough: Improving User Engagement and Satisfaction Through Automatic Quality Evaluation. In *SIGIR '15: Industry Track*, 2015.

[9] E. M. Rodrigues and N. Milic-Frayling. Socializing or knowledge sharing?: characterizing social intent in community question answering. *CIKM*, 2009.

[10] A. Shtok, G. Dror, Y. Maarek, and I. Szpektor. Learning from the past: answering new questions with past answers. In *WWW '12: Proceedings of the 21st international conference on World Wide Web*, 2012.

[11] Z. Zhao and Q. Mei. Questions about Questions: An Empirical Analysis of Information Needs on Twitter. In *WWW*, 2013.