Finding Informative Q&As on Twitter

Kanghak Kim*
KAIST, South Korea
kanghak.kim@kaist.ac.kr

Sunho Lee *
South Korea
sunholeee@gmail.com

Jeonghoon Son *
South Korea
jsonarie@gmail.com

Meeyoung Cha
KAIST, South Korea
meeyoungcha@kaist.ac.kr

ABSTRACT

Question & Answer (Q&A) behaviors on social media have huge potential as a rich source of information and knowledge online. However, little is known about how much diversity there exists in the topics covered in such Q&As and whether unstructured social media data can be made searchable. This paper seeks the feasibility of utilizing social media data for developing a Q&A service by examining the topic coverage in Twitter conversations. We propose a new framework to automatically extract informative Q&A content using machine learning techniques.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H3.1 Content Analysis and Indexing;

General Terms

Human Factors, Measurement, Experimentation

Keywords

Social Media; Community Question Answering; Twitter

1. INTRODUCTION

Community Question Answering (CQA) services such as Yahoo! Answers (http://answers.yahoo.com) and Naver's Knowledge-IN (http://kin.naver.com) have long been an important source for finding answers to user-generated questions. The basic idea of CQA services is to provide a platform that connects askers to possible answerers and to develop a reputation system that enhances the quality of answers. However, one of the critical factors for the success of CQA services is their searchability. By making the content searchable, CQA has become publicly useful.

With the rapid growth of CQA services, however, the responsiveness and the quality of answers have been challenged. Answering behaviors in CQA usually follows a power-law distribution, in which a disproportionately small number of active users account for most of the answers on the sites. Due to this limitation in incurring wide participation, the ratio of answered questions has decreased and the credibility of the answered Q&A content has degraded, as CQA services have grown [5, 1].

The challenges facing CQA services have recently shed light on social media-based Q&A. Recently, researchers have found that social media users utilize their social networks as a place for asking questions [2, 4, 3]. The characteristics of question answering behaviors in social networks have advantages against the existing CQA sites in terms of credibility and responsiveness. While most of the askers receive answers from strangers in CQA, social media users receive answers from their trustworthy friends in the network. In addition, the questions are distributed across the asker's entire social network, allowing everyone in the network to contribute, which makes the system more responsive.

Despite the potential benefits of utilizing social media data, it is unclear whether Q&A conversations on social media can be mined effectively as a stand-alone service. It is important to know whether (1) social media data cover a wide range of topics as existing CQA services and whether (2) there is a framework to extract high-quality Q&A content from the massive amount of unstructured and heterogeneous social media data. In this paper, we test the feasibility of turning social media into a valuable Q&A platform by addressing these challenges based on a large amount of data gathered from the Twitter microblog network.

2. METHODOLOGY

We collected tweets from September 1st to 30th in 2013 using the Twitter Streaming API with the parameter value set to the question mark and preset question bigrams. To avoid any bias in the data collection process caused by the posting time, we randomly sampled postings from the collected *status id* and finally gathered 22,118 postings and the corresponding answers. Then, we evaluated the informativeness of each Q&A conversation with a group of elite workers in Amazon Mechanical Turk, assuring the quality of answers with a set of gold standard answers. Of the collected Q&A tweets, 3,570 (17%) of the Q&As were evaluated to be 'informative' by the workers in the Mechanical Turk. Among the data, we used 70% of our data as our training set, and the remaining data as a test set.

*This research was done while the marked authors worked at WIV LABS, South Korea.

WWW'14 Companion, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2577291.

Table 1: List of features and their descriptions

Category	Feature
Text	Term importance (information Gain)
	Readability (POS probability)
	Slang ratio
	Length of conversation
	Num of participants in Q&A
	Pronoun ratio
	Unique token ratio
User	Num of profile images with faces
	PageRank of the asker and answerers
Context	Word similarity of non-stop words
	Topic similarity (LDA - JS Divergence)
	Prob. of question to be information-seeking
	Words of appreciation from the asker

3. RESULTS

We present results on the topic coverage of Twitter Q&A conversations and searchability of informative Q&A content.

3.1 Topic Coverage

In order to test whether Twitter-based Q&A conversations cover a wide range of topics, we need a comparable guideline. We collected 205,609 Yahoo! Answers items posted during the same period as those for our Twitter datasets. We then randomly sampled the equal amount of Q&A content from Yahoo! Answers and the Twitter data. We merged two datasets and executed Latent Dirichlet Allocation on the merged set with the number of topics set to 20. We then manually labeled each topic by observing the highly probable terms in the topic and excluded topics that were not clearly defined. Finally, we assigned labels to 12 topics, and counted the number of postings in each topic on Twitter Q&A and Yahoo! Answers, respectively.

Our analysis demonstrates that the overall probability distributions of informative Q&As on Twitter cover a wide range of topics as the well-established CQA service. Furthermore, Twitter-based content even showed competitive advantage to Yahoo! Answers on several topics like *entertainment, local information, music, technology.*

3.2 Informative Q&As

The features are categorized into 3 groups in Table 1: text, user, and context-related. To train the model, we used Support Vector Machine, Neural Network, and Random Forest (RF). Table 2 shows the performances of the best performing classifier (RF), varying the subset of features. Among the set of features examined, the *prob. of the question to be informative-seeking, slang ratio, pronoun ratio, unique token ratio* and *topic similarity* are shown to be powerful predictors according to the *Mean Decrease Accuracy* measure. The result shows that our proposed classifier successfully classifies informative Q&As from the overall contents. However, it is still unclear whether Q&As on Twitter can be useful outside Twitter. Table 3 presents the examples of classified Q&As to help understand the usefulness.

4. CONCLUDING REMARK

We conclude that unstructured data within social media is promising for finding answers to information-seeking questions. We demonstrated this feasibility based on the fact that Twitter conversations on Q&As involve a wide range of topics and they could also be identified automatically. Moreover, the limited length of postings on Twitter might be a strong point in its favor when it is applied to a mobile-centered search service, where people prefer shortened summaries to long articles.

Table 2: Performance of the proposed classifier

Features	Precision	Recall	AUC
Text(T)	0.651	0.708	0.712
Context(C)	0.622	0.645	0.688
User(U)	0.588	0.572	0.619
T+C	0.696	**0.720**	0.765
T+U	0.651	0.703	0.717
T+C+U	**0.702**	0.708	**0.770**

Table 3: Examples of classified Q&As and their prediction scores

Q&A contents		Pred
A:	Anybody know how to transfer pictures from a iPhone to a galaxy s4?	0.90 (info)
B:	@A You can do it via Dropbox app or email. You can also back up pics to PC first via iTunes or iTransfer: http://bit.ly/NfMcO2	
A:	Where is the best place for a brunch in London? #brunch #saturday #london...	0.67 (info)
B:	@A I LOVE Kopapa on Monmouth Street, Covent Garden x	
A:	Is anyone selling an iPhone or galaxy or know someone selling for sprint?	0.18 (non-info)
B:	@A I'm selling my iPhone	
A:	@B I just dm'd you my number shawty	

Acknowledgements

The authors would like to thank Dongug Kim and Changyu Jang for making the data available and their valuable comments. Meeyoung Cha was funded by the Basic Science Research Program through the National Research Foundation of Korea funded by the Ministry of Science, ICT & Future Planning (2011-0012988).

5. REFERENCES

[1] Y. Liu, J. Bian, and E. Agichtein. Predicting information seeker satisfaction in community question answering. In *Proc. of the SIGIR*, 2008.

[2] 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 *Proc. of the ACM CHI*, 2010.

[3] J. Nichols and J.-H. Kang. Asking questions of targeted strangers on social networks. In *Proc. of the ACM CSCW*, 2012.

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

[5] A. Shtok, G. Dror, Y. Maarek, and I. Szpektor. Learning from the past: answering new questions with past answers. In *Proc. of the WWW*, 2012.