

Social Information Retrieval for Online Community Question Answering Services

Jun Choi Lee¹, Yu-N Cheah², and Wai Shiang Cheah³

¹Universiti Malaysia Sarawak, Malaysia, {jclee; wscheach}@unimas.my

²Universiti Sains Malaysia, Malaysia, yncheah@usm.my

ABSTRACT

Online Community Question Answering Services is a platform that allows users to seek for answers anytime and anywhere. The services cover a large collection of unstructured knowledge and become a knowledge repository for problem solving. The answers in the services are generated by other public user. This environment creates a large collection of unstructured knowledge, which can be use as knowledge source for other intelligent processes. From the review, it is a challenge to access accurate information from the Online Community Question Answering services. This study propose a social information retrieval method to access the knowledge from the services. The proposed method query and retrieve information the online services through query expansion. Then, the retrieved information is filtered through a similarity analysis between a question with a query statement and semantic relatedness between answers and the query statement. The accuracy of the proposed approach is evaluated through human evaluation.

Keywords: Social Information Retrieval, Online Community Question Answering Services, Semantic Relatedness.

I INTRODUCTION

Online Community Question Answering Services (CQA) like Yahoo! Answers, Quora and Baidu Knows are online platforms to guide users in seeking for answers. A question is posted online. Then, people can post answers based on their expertise and personal experience. Lots of studies are conducted based on online CQA. These studies ranged from artificial intelligent to human behavior studies. However, most of these studies are conducted based on pre-collected data from online CQA data. We argue that the pre-collected CQA data is not suitable for real-life applications. This is because the pre-collected data is only a snapshot of the online CQA. Application of pre-collected data in real life application will require the pre-collected data to be updated from time-to-time. Therefore, knowledge in pre-collected data cannot represent the overall knowledge in online CQA.

On the other hand, it is a time consuming and complex process in working on pre-collected data. Besides limitation of knowledge, application of pre-collected CQA data in real-life applications also required the data to be stored in a structured storage which allows it to be query and retrieved. This is a resource consuming task as online CQA constantly receive huge amount of user generated data daily.

Social Information Retrieval (SIR) is the study to apply social media in information retrieval. The process can be conducted by either query and retrieve information from social media, or applying social profile data in performing information retrieval. In this study, a social information retrieval framework is proposed to query and retrieve relevant information from selected online community question answering service.

This study presents a Social Information Retrieval to query and retrieve relevant information from the selected online CQA. Section two explains the background and motivation of the proposed SIR, while section three provides a detailed description of the proposed SIR. The evaluation of the proposed SIR and its result are presented in the section four. The paper is concluded in Section five.

II BACKGROUND AND MOTIVATION

Information Retrieval (IR) is science to gain access and retrieve relevant information from different information sources in order to satisfy information thirst of a user. This definition is shared by Salton (1968), Baeza-Yates and Ribeiro-Neto (2010), and Bouadjenek, Hacid and Bouzeghoub (2015). Boudjenek (2013) defined SIR as “The process of leveraging social information (both social relationships and the social content), extracted from social platforms, to perform an IR task with an objective of better meeting users’ needs.” Goh and Foo (2007) introduced a platform to share knowledge through Social Information Retrieval. Bouadjenek, Hacid and Bouzeghoub (2015) reported on IR approaches and platforms for social networks. From the review, most of the studies on SIR were on expanding the IR model with social network applications. These studies provide fundamentals for involving data from social network in various stages in IR architecture.

CQA also provide opportunities in various computer science and information science research. Researches in CQA covers question classification, answer quality evaluation, experts routing, user behavior analysis and answers generations. Shtok, Dror, Maarek and Szpektor (2012) uses CQA as evaluation platform for their proposed Question Answering (QA) system. To date, most of the CQA related researches are based on question answer archives. The archives are good in evaluating the proposed models, but it has some limitations in real-life applications. The limitations include the need for a huge storage capacity to store the archives, and having not up-to-date data in the archives.

Yahoo Answers provided Application Program Interface (API) to access certain information from their online CQA repository. This service was adopted by Chen (2009) and Kai (2011) in their QA research. However, the service only available for a short period of time before it was discontinued in mid of year 2014. Therefore, there is a need to look for an alternative to access directly the online CQA repository for the relevant third party research to forward from theoretical stage to actual implementation.

A good SIR for CQA should able to retrieve useful information from their sources. Since responses in CQA are user generated, the quality of the answers is varying. Therefore it is necessary to identify quality answers from the data retrieved from the CQA. Answer quality evaluation is a research domain in IR studies. Jeon, Croft, Lee and Park (2006) attempted to predict the quality of answers 13-nontextual features in Maximum entropy learning model from the finding, they discovered that 1/3 of the dataset from Naver.com have a quality issue, and 10% of the total dataset are identified with bad answers. Jeon, Kim and Chen (2010) extended the study using price as a factor to identify quality answers in fee-based CQAs. Harper, Raban, Rafaeli and Konstan (2008) presented the differences between quality answers in different CQAs and observed how users accept the different quality criterion. Liu et al. (2008) tried to predict the answer quality using non-contextual information. Fichman (2011) studied the quality of the answers from of different online Community QA Services. Agichtein, Castillo, Donato, Gionis and Mishne (2008) applied feature extraction technique to determine a quality of the answers given. Adamic, Zhang, Bakshy, and Ackerman (2008) explored a method to predict answer quality using question and answer provided. Answer Quality prediction study done by Shah and Pomerantz (2010) were based on the information extracted from questions, answers and user profile. Blooma, Chua and Goh (2010) also explore the methods to select best answers in CQA. Tian, Zhang and Li (2013) investigated different

features from question and answers in the CQA, www.stackoverflow.com, to measure answer quality. Their study shows that the most significant factors in determining the answers quality were the number of responses and the minimum similarities between the answers.

In the previous studies, the most effective methods to evaluate the answer quality is based on non-textual features like length of answers and user reputation. However, the actual answer quality should still base on the content of the answer.

This study aims to overcome two limitations in retrieving quality information from CQA by introducing a new SIR for CQA. In the proposed SIR, a mechanism to access the online CQA repository is proposed to overcome the limitation of using archives. This study also proposed to filter quality information from the online data using text semantic relatedness. This method aims to measure the quality of the answers or data from CQAs based on content, rather than non-contextual features.

III PROPOSED SOCIAL INFORMATION RETRIEVAL

The proposed social information retrieval for online community question answering services consists of two major components: query and accessing selected online community question answering services, and filtering useful information based on the query posted by a user.

A. Query And Access Selected Online CQA

Accessing the information from online CQA is one of the most crucial and challenging components in the proposed SIR. Most online CQAs do not have API that allows user to access the collection of question and answers in their services, therefore this is the major challenge in information retrieval for these services.

Every online CQAs have at least one web services that allows user to access the question and answers in their web portal. For example, Yahoo! Answers provides Rich Site Summary (RSS) feeds that allows user to access a snapshot of the online CQA, while other online CQA like Quora and Baidu allows user to search the content of the online CQA through their web addresses. Although these web services only provides a glimpse of information in the selected CQAs, but it is enough to function as a gateway to retrieve information from these online CQAs. Therefore, the proposed SIR exploited these web services to access and query a selected online CQAs. This task is done by applying query expansion on the RSS or web link uses to retrieve information on the selected CQAs.

The query expansion is done by placing the user query in the web service links. The results of this query expansion is a list of web links. Each link represents a dedicated webpage with a specific questions and answers from the selected online CQA. The proposed SIR then will harvest the Hyper Text Markup Language (HTML) data from each of this links using web access. These HTML data contains all the information regarding a particular questions and its answer. These information are later extracted through pattern recognition process. The pattern used in extracting question and answers information from HTML data are vary depends on the CQAs, this is because each CQA have their own web design.

The final results for this process is a list of questions and their answers that related to the query statement provided by a user. However, these information need to be further filtered to obtain useful information.

B. Filtering Relevant Information

The questions and answers extracted through the query expansion and pattern recognition in previous section are user generated information. These data consists some noises that are not important such as user conversations, emotional arguments or even advertisement spam. In this process, the proposed SIR applied text semantic relatedness measures to filter the important information related from the collected data. In the study conducted by Lee and Cheah (2015), Text semantic relatedness that used to measure the degree of relationship between two texts has reasonable accuracy in identify relevant information in online CQAs. In that study, Lee and Cheah applied a text semantic relatedness that is based on Wul & Palmer (WUP) measures (Wu & Palmer, 1994) in WordNet to identify relevant answers from Yahoo! Answers.

C. Output of Proposed SIR

The result returned by the proposed SIR Framework is a list of text or answer(s) that being filtered. However the result is in the form of answers. The content of the information is highly dependent on the query statement provided by user. The results also shows another characteristic, where the answers in the results are highly redundant. This happen because of these answers are extracted from questions that share high degree of similarity. The proposed SIR remains the high redundancy characteristic of the result for it may be useful in many application of the SIR results such as in Decision Supports System or Question Answering system.

IV IMPLEMENTING THE PROPOSED SIR ON YAHOO! ANSWERS

In the final section of this study, the proposed SIR is evaluated through an implementation of the proposed SIR. For the implementation, this study had selected

Yahoo! Answers to serve as knowledge source for the proposed SIR. Yahoo! Answers is selected for implementation in this study because it is one of the most popular and active online CQAs currently available on Internet. The study from Harper, Raban, Rafaeli and Konstan (2008) had indicated that Yahoo! Answers not only have more responses from user compare to other online CQAs, the responses in Yahoo! Answers are also higher quality compare to other online CQAs. Another crucial element for selecting Yahoo! Answers as the online CQA for the proposed SIR is because Yahoo! Answers is open question answering platform, where there is no topic limitation in posting the question.

The querying of Yahoo! Answers online repository is done through RSS query expansion. Yahoo! Answers provides Rich Site Summary (RSS) feed for user to access a snapshot of repository in Yahoo! Answers. The RSS feature also provides search function, where user can directly search the Online Yahoo! Answers repository. The social information retrieval access Yahoo! Answers repository by reconstruct the RSS feed based on the user query. The RSS feed for the reconstructed link is obtained through normal HTTP web request. The RSS feed returns from Yahoo! Answers are (Extensible Markup Language) XML that consists of questions and their Yahoo! Answers links. No answer is provided in the RSS feed.

The RSS feed only consists a numbers of questions and some basic information such as the question, the link to the question in Yahoo! Answers, the date the question is posted and some brief description regarding the question posted by the author. These information need to go through several process before it can be extracted for further use.

Each link in the RSS feed represents a dedicated webpage in Yahoo! Answers for a particular question. The raw format of the webpage (HTML) is obtained through a web request for each of the question in the RSS feed. Before extracting the information from the webpage, all HTML tags in the document is removed. The remaining text is an array of text. The answers for each question are extracted from this array of text using a text pattern rules. The text pattern rules used in extracting the information is shown in Table 1. These rules are obtained through detailed study on the text pattern exists in the Yahoo! Answers webpage. There are three type of possible answers in a page: "Best Answer", "First Answer" and "Other Answers".

Table 1. Text Pattern To Extract Information From Webpage After Removing HTML Tag

Information	Text Patterns
Best Answer	<p>Start: After the line starts with “Best Answer:”</p> <p>End: before the line starts with “Source(s)”</p> <p>AND</p> <p>The third line starts with “&middot” and ends with “ago”</p>
First Answer	<p>Start: the line starts from “Oldest”</p> <p>End: before the line starts with “Source(s)”</p> <p>AND</p> <p>The third line starts with “&middot” and ends with “ago”</p>
Other answer	<p>Start: After the line starts with “Report Abuse”</p> <p>End: before the line starts with “Source(s)” AND</p> <p>The third line starts with “&middot” and ends with “ago”</p>

All the possible answers for a question in the question and answer page are extracted for further usage. However, as mentioned in Section III, the answers collected need to be filtered to obtain relevant information regarding the query statement. And the filtering is conducted through the use of text semantic relatedness based on WUP measure in WordNet, with a threshold value of 0.6. The threshold value used to filter the answer is based on the experiment conducted by Lee and Cheah (2015). The filtering result is the final output of the proposed SIR. Figure 1.0 shows the screenshot of an implementation of the proposed SIR.



Figure 1. Screenshot Of The Implementation For The Proposed SIR

V CONCLUSION

This study presented a SIR framework to extract and filter information from a selected CQAs. The SIR queries selected CQA through web request and harvest questions and answers. Then, useful information from the question answer is obtained

through a filtering process. The filtering process uses a text semantic relatedness in order to measure relationship between a query statement and candidate answers. The final output of the SIR is a list of answers from related CQAs. These outputs can serve as information for further knowledge process. An implementation of the SIR based on Yahoo! Answers as knowledge source is developed in this study. This implementation demonstrates the capability of the proposed SIR in querying an existing online CQAs and return a set of answers that presumable relevance to the query statements.

The proposed SIR have a wide range of potential applications. It can be applied as knowledge source for various intelligent process such as decision support systems or answer generation system. The proposed SIR is also useful in studies such as event summarizer and human behavior studies.

Although the proposed SIR currently is able to retrieve relevant information from selected CQAs. The study can be furthered in various ways. One of the future enhancement for this study is to incorporate more than one social platform to retrieve information relevant to the query statement. The current implementation of the SIR is only based on Yahoo! Answers, but it had been identified that other CQAs such as Quora, and Baidu Knows is also compatible to serve as knowledge source in the proposed CQAs. Combining different CQAs in the proposed SIR will enrich the knowledge pool to obtain information. Besides CQAs, the proposed SIR framework should also expands to other social platform such as Facebook and Twitters to further enrich the content for the information retrieval process.

Besides enriching the content of the results through expand the knowledge source to various social platforms. The proposed SIR can be further enhanced by using a more comprehensive filtering process. The current filtering processes is done using a text semantic relatedness that based on WUP measure in WordNet. This semantic relatedness still possess some limitation due to the limited real-world concept knowledge in WordNet. Therefore, filtering the information in SIR using a more comprehensive filtering technique such as using Latent Semantic Analysis (LSA) will improve the quality of the information retrieved using the SIR framework.

Finally, this study only demonstrated the proposed SIR through a simple implementation. A detailed evaluation is needed to understand the performance of the proposed SIR in information retrieval.

ACKNOWLEDGMENT

This study is supported by Universiti Malaysia Sarawak (UNIMAS) and Universiti Sains Malaysia (USM).

REFERENCES

- Adamic, L. A., Zhang, J., Bakshy, E., & Ackerman, M. S. (2008). Knowledge sharing and yahoo answers: everyone knows something. In Proceedings of the 17th international conference on World Wide Web (pp. 665-674). ACM.
- Agichtein, E., Castillo, C., Donato, D., Gionis, A., & Mishne, G. (2008). Finding high-quality content in social media. In Proceedings of the 2008 International Conference on Web Search and Data Mining (pp. 183-194). ACM.
- Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern IR* (Vol. 463). New York: ACM press.
- Blooma, M. J., Chua, A. Y. K., & Goh, D. H. L. (2010). Selection of the best answer in CQA services. In Seventh International Conference on Information Technology: New Generations (ITNG), 2010 (pp. 534-539). IEEE.
- Budan, I. A., & Graeme, H. (2006). Evaluating WordNet-Based Measures of Semantic Distance. *Computational Linguistics*, 32(1), 13-47.
- Bouadjeneq, M. R., Hacid, H., & Bouzeghoub, M. (2015). Social Networks and IR, How Are They Converging? A Survey, a Taxonomy and an Analysis of Social IR Approaches and Platforms. *Information Systems*.
- Bouadjeneq, M. R. (2013). *Infrastructure and Algorithms for Information Retrieval Based On Social Network Analysis/Mining* (Doctoral dissertation, Versailles-Saint-Quentin-Yvelines).
- Chen, L. (2009). *Recommending best answer in a collaborative QA system*. (Master dissertation, Queensland University of Technology)
- Fichman, P. (2011). A comparative assessment of answer quality on four QA sites. *Journal of Information Science*, 37(5), 476-486.
- Goh, D., Foo, S. (2007). *Social IR Systems: Emerging Technologies and Applications for Searching the Web Effectively: Emerging Technologies and Applications for Searching the Web Effectively*. IGI Global.
- Harper, F. M., Raban, D., Rafaeli, S., & Konstan, J. A. (2008). Predictors of answer quality in online Q&A sites. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 865-874). ACM.
- Jeon, G. Y., Kim, Y. M., & Chen, Y. (2010). Re-examining price as a predictor of answer quality in an online Q&A site. In Proceedings of the SIGCHI conference on human factors in computing systems (pp. 325-328). ACM.
- Jeon, J., Croft, W. B., Lee, J. H., & Park, S. (2006, August). A framework to predict the quality of answers with non-textual features. In Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 228-235). ACM.
- Kai, W. (2011). *Retrieving questions and answers in community-based Question Answering services* (Doctoral dissertation, National University of Singapore).
- Lee, J.C. & Cheah Y-N. (2015b). Identifying Relevant Answers in Online CQA Services. In Conference of Information Technology in Asia 2015 (CITA'15). 4th – 5th August, 2015. Kuching, Malaysia.
- Liu, Y., Li, S., Cao, Y., Lin, C. Y., Han, D., & Yu, Y. (2008). Understanding and summarizing answers in community-based QA services. In Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1 (pp. 497-504). Association for Computational Linguistics.
- Salton, G. (1968). Automatic information organization and retrieval. In *Information Processing & Management*, 24(5): 513-523.
- Shah, C., & Pomerantz, J. (2010). Evaluating and predicting answer quality in community QA. In Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval (pp. 411-418). ACM.
- Shtok, A., Dror, G., Maarek, Y., & Szpektor, I. (2012). Learning from the past: answering new questions with past answers. In Proceedings of the 21st international conference on World Wide Web (pp. 759-768). ACM.
- Tian, Q., Zhang, P., & Li, B. (2013). Towards Predicting the Best Answers in Community-based Question-Answering Services. In ICWSM.
- Wu, Z., & Palmer, M. (1994). Verbs semantics and lexical selection. In Proceedings of the 32nd annual meeting on Association for Computational Linguistics (pp. 133-138). Association for Computational Linguistics.