Document Summarization For Answering Non-Factoid Queries

Sheik John , Miss G.Keerthana, Sri.V.Bhaskara Murthy
MCA Student, Assistant Professor, Associate Professor
Dept Of MCA
B.V.Raju College, Bhimavaram

ABSTRACT

We formulate a document summarization method to extract passage-level answers for non-factoid queries, referred as answer-biased summaries. We propose to use external information from related Community Question Answering (CQA) content to better identify answer bearing sentences. Three optimization-based methods are proposed: (i) query-biased; (ii) CQA-answer-biased; and (iii) expanded-query-biased, where expansion terms were derived from related CQA content. A learning-to-rank-based method is also proposed that incorporates features extracted from related CQA content. Our results show that even if a CQA answer does not contain a perfect answer to a query, their content can be exploited to improve the extraction of answer-biased summaries from other corpora. The quality of CQA content is found to impact on the accuracy of optimization-based summaries, though medium quality answers enable the system to achieve a comparable (and in some cases superior) accuracy to state-of-the-art techniques. The learning-to-rank-based summaries, on the other hand, are not significantly influenced by CQA quality. We provide a recommendation of the best use of our proposed approaches in regard to the availability of different quality levels of related CQA content. As a further investigation, the reliability of our approaches was tested on another publicly available dataset.

II. INTRODUCTION

Current search engines usually present single direct an-swers on a search result page for some popular factoid queries (e.g. current weather) [1], and for some entity queries. Some major search engines have also started to present single passages (so-called featured snippets) in response to more verbose informational queries. Accord-ing to the Moz SERP features tracker1, in 2017, these pas-sages appears in 15% of queries submitted to Google, though with some errors Such direct answers can improve a user's search expe-rience [1]–[3]. They may also lead to good abandonment [4], where users find what they need in the result page and therefore do not need to read the full document. By re-moving the document reading step, user time can sub-stantially be saved. This is as reported by Smucker and Clarke [5] that users spent 67% of their searching time reading webpages. Direct answers provide the most bene-fit to users who search on devices with limited screen size and low bandwidth (e.g. mobile search) as clicking through can incur additional costs at the user's end.

While non-factoid queries are the most frequently asked questions on the web [6], [7], research on finding answers for this type of query has not been extensively explored. Some past work was conducted to generate passage-level answers to a non-factoid query [3], [8]–[10]. However these approaches do not explore the idea of us-ing automatic summarization. We argue that summarization techniques can be

beneficial in tackling this problem because answers to non-factoid queries may consist of a number of sentences scattered in the underlying docu-ment (potentially with some overlap in the content) [9].

II. EXISTING SYSTEM

- Question Answering (QA) is an information retrieval task that returns answers in response to natural language questions. Commonly supported question types in this research include factoid, list, and definition questions [14]. Answering different types of questions generally relies on different techniques. Previously, much of the attention in the research of question answering has fo-cused on answering factoid and list questions, which are the main themes of the TREC QA track [15].
- Our work is different from TREC QA as we focus on non-factoid questions, such as the ones from TREC Tera-byte topics: "What allegations have been made about Enron's culpability in the California Energy crisis?", in which it may not be satisfied with just one or a list of factoids. A recent method in factoid QA that has superior performance on TREC QA track data has also been shown to perform poorly for these kinds of questions [16].

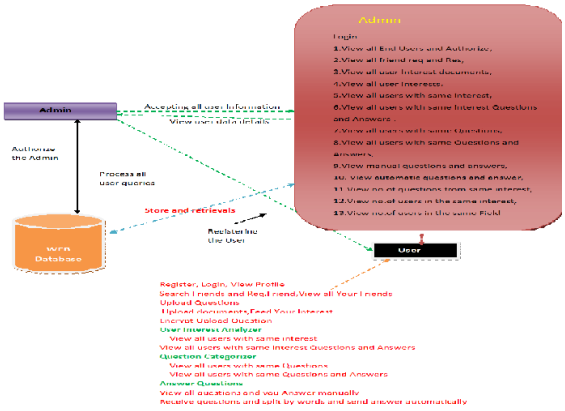
- Community Question Answering (CQA) is a service that allows users to post questions and elicit answers from other peers. Major CQA websites, such as Yahoo! An-swers, Quora, and Stack Overflow, continue to see a growing user base. It is reported that Yahoo! Answers (YA) attracted 7,000 questions and 21,000 answers every hour in 2012.3 This sheer amount of data has attracted a lot of research activities: predicting answer quality in CQA [17]; predicting the satisfaction of the original ques-tion asker [18] and web searcher [19] with CQA answers; answering factoid [20] and how-to web queries [8]; en-hancing document summaries [21]; and summarizing CQA answers [22], [23].

- Finding answers for non-factoid queries remains a critical challenge in web question answering, and one difficulty is the vocabulary mismatch between questions and an-swers. Keikha et al. [9] has shown that state-of-the-art passage retrieval methods that focus on topical relevance are not effective for this task. A recent forum that is relat-ed to answering non-factoid questions (that came from real YA users) is the TREC LiveQA track [24]. The quality of the best performing run in this track is shown still far from human level, indicating the complexity of the task.

III. PROPOSED SYSTEM

- The system proposes a novel use of CQA content in a summari-zation algorithm for locating answer-bearing sentences in the document.
- The system proposes three optimization-based methods and a learning-to-rank-based method for answering non-factoid queries. These methods are empirically evaluated against state-of-the-art techniques.
- The system analyses the effect of quality of related CQA content on our proposed methods. Then, we give recommenda-tions on the best use of our methods in regard to the availability of different quality levels of CQA answers.

IV. ARCHITECTURE DIAGRAM



V. MODULES

User Interest Analyzer:

User Interest Analyzer utilizes each user's profile information in the social network and user interactions (answers provided and questions asked) to determine the interests of the user in the predefined interest categories. This is because if a user asks or answers questions in an interest category, (s)he is likely to be interested in this particular category.

Question Categorizer:

The primary task of Question Categorizer is to categorize a question into predefined interest categories based on the topic(s) of the question. We also allow users to input selfdefined tags associate with questions, which are analyzed in question parsing. Question Categorizer generates a vector of question Qi's interests, denoted by VQi , using a similar algorithm While processing a question, SocialQ&A uses WordNet to examine the tags and text of the question and generates a token string. The tokens are compared to SocialQ&A's Synset to determine the categories where the question belongs. We have calculated the interest weight without normalization in order to predict the user intelligence to answer a question of Interest.

Question-User Mapper:

Question-User Mapper identifies the appropriate answerers for a given question. The potential answer providers are chosen from the asker's friends in the online social network. Note that the changes in a user's friends in the online social network do not affect the performance of SocialQ&A as it always uses a user's current friends. To check the appropriateness of a friend (Uk) as an answer provider for a question, two parameters are considered: i) the interest similarity between the interest vectors of the friend and the question (denoted by I;Uk); and ii) the social closeness between the friend and the asker (denoted by C;Uk). The former represents the potential capability of a friend to answer the question, and the latter represents the willingness of a friend to answer the question.

VI. CONCLUSION

We propose to use external information from related CQA content to guide the extraction of an answer-biased summary from each retrieved document. Three optimiza-tion-based methods and a learning-to-rank-based method were proposed. Our results show that the related CQA content, that do not necessarily contain perfect answer to the query, are useful to extract better answer-biased summaries from documents. This answers RQ1. The quality of CQA content is shown to have significant effect to the accuracy of optimization-based summaries. In con-trast, the significant effect of CQA quality is not found on the accuracy of learning-to-rank-based summaries. The learning-to-rank-based method consistently performs well on different level of CQA quality. This answers RQ2.

REFERENCES

[1] L. B. Chilton and J. Teevan, "Addressing People's Information Needs Directly in a Web Search Result Page," in Proc. 20th Int. Conf. World Wide Web, 2011, pp. 27–36.

[2] D. Lagun, C.-H. Hsieh, D. Webster, and V. Navalpakkam, "To-wards Better Measurement

of Attention and Satisfaction in Mo-bile Search," in Proc. 37th Annu. Int. ACM SIGIR Conf. Res. Devel-op. Inf. Retrieval, 2014, pp. 113–122.
[3] M. S. Bernstein, J. Teevan, S. Dumais, D. Liebling, and E. Horvitz, "Direct Answers for Search Queries in the Long Tail," in Proc. SIGCHI Conf. Human Factors Comput. Syst., 2012, pp. 237–246.
[4] J. Li, S. Huffman, and A. Tokuda, "Good abandonment in mobile and PC internet search," in Proc. 32nd Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2009, pp. 43–50.
[5] M. D. Smucker and C. L. Clarke, "Time-based calibration of effec-tiveness measures," in Proc. 35th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2012, pp. 95–104.
[6] V. Jijkoun and M. de Rijke, "Retrieving answers from frequently asked questions pages on the web," in Proc. 14th ACM Int. Conf. Inf. Knowl. Manage., 2005, pp. 76–83.
[7] T. Nguyen et al., "MS MARCO: A Human Generated MAchine Reading COmprehension Dataset," in Proc. NIPS Workshop, 2016.
[8] I. Weber, A. Ukkonen, and A. Gionis, "Answers, Not Links: Ex-tracting Tips from Yahoo! Answers to Address How-to Web Queries," in Proc. 5th ACM Int. Conf. Web Search Data Mining, 2012, pp. 613–622.
[9] M. Keikha, J. H. Park, W. B. Croft, and M. Sanderson, "Retrieving Passages and Finding Answers," in Proc. Australasian Document Computing Symposium, 2014, pp. 81–84.
[10] R. Soricut and E. Brill, "Automatic Question Answering Using the Web: Beyond the Factoid," Inf. Retr., vol. 9, no. 2, pp. 191–206, Mar. 2006.
[11] E. Cutrell and Z. Guan, "What are you looking for?: an eye-tracking study of information usage in web search," in Proc. SIGCHI Conf. Human Factors Comput. Syst., 2007, pp. 407–416.
[12] K. Williams, J. Kiseleva, A. C. Crook, I. Zitouni, A. H. Awadallah, and M. Khabsa, "Detecting good abandonment in mobile search," in Proc. 25th Int. Conf. World Wide Web, 2016, pp. 495–505.
[13] Q. Liu, E. Agichtein, G. Dror, Y. Maarek, and I. Szpektor, "When web search fails, searchers become askers: understanding the transition," in Proc. 35th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2012, pp. 801–810.
[14] O. Kolomiyets and M.-F. Moens, "A survey on question answer-ing technology from an information retrieval perspective," Inf. Sci., vol. 181, no. 24, pp. 5412–5434, Dec. 2011.
[15] H. T. Dang, D. Kelly, and J. J. Lin, "Overview of the TREC 2007 Question Answering Track.," in Proc. TREC, 2007, vol. 7, p. 63.
[16] L. Yang et al., "Beyond Factoid QA: Effective Methods for Non-factoid Answer Sentence Retrieval," in Adv. Inf. Retrieval: 38th Eu-ropean Conf. Inf. Retrieval, 2016, pp. 115–128.
[17] C. Shah and J. Pomerantz, "Evaluating and predicting answer quality in community QA," in Proc. 33rd Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2010, pp. 411–418.
[18] Y. Liu, J. Bian, and E. Agichtein, "Predicting information seeker satisfaction in community question answering," in Proc. 31st Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2008, pp. 483–490.
[19] Q. Liu et al., "Predicting Web Searcher Satisfaction with Existing Community-based Answers," in Proc. 34th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2011, pp. 415–424.
[20] J. Bian, Y. Liu, E. Agichtein, and H. Zha, "Finding the right facts in the crowd: factoid question answering over social media," in Proc. 17th Int. Conf. World Wide Web, 2008, pp. 467–476.