Expert Systems With Applications 44 (2016) 92–101

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

High quality information extraction and query-oriented summarization for automatic query-reply in social network

Min Peng[a,c,1], Binlong Gao[b,2], Jiahui Zhu[a,3], Jiajia Huang[a,4], Mengting Yuan[a,5,*], Fei Li[a,6]

[a] *Computer School of Wuhan University, No. 299, Bayi Road, Wuchang District, Wuhan, China*
[b] *Netease Research Hangzhou, No. 599,Wangshang Road, Binjiang District, Hangzhou, China*
[c] *Shenzhen Institute of Wuhan University, Yuexing Road, Nanshan District, Shenzhen, China*

ARTICLE INFO

Keywords:
Query-reply
Social network
Information extraction
Summarization

ABSTRACT

In this paper, we propose a new method for automatic query-reply in social network. Information extraction and query-oriented summarization method are applied here to reply people's query. There are few effective and commonly used methods on filtering the redundancy and noise of the raw data, which results in the poor quality of the reply. Due to the characteristics of social network messages, we pay more attention to reducing the noise and eliminating the redundancy of the messages to ensure the quality of the final reply. First, we propose an information extraction method to extract high quality information from social network messages, which is based on time-frequency transformation. Second, query-oriented text summarization is implemented to generate a brief and concise summary as the final reply, which is based on the scoring, ranking and selection of sentences of high quality social network messages produced by previous step. Experimental results show that the research is effective in filtering the redundancy and noise of social network messages, the final query-reply results outperform other commonly used methods' results in both automatic evaluation and manual evaluation. Through our approach, noise and redundancy of social network messages are effectively filtered. Certainly, our method improves the quality of the reply for people's query.

1. Introduction

Nowadays, as a broadcast medium for broadcasting short messages, social network has become popular rapidly. We live amidst seas of short text documents in social network where the amount of social network users and messages is ever-increasing. For example, *Twitter* has attracted more than 1000 million registered users. The average number of daily active users has reached 284 million, and the average number of daily posted tweets has reached 400 million. Social network is available for users to get and report novel information or express themselves, because the data of it cover a broad range of events or topics and people can broadcast information in a live manner.

People tend to get precise and brief answer when they put forward a question or retrieve by the search engine of social network. Although manual reply provides precise answers, the cost of this approach is unacceptable as there are numerous users and questions in the social network. Automatic reply is a better choice because of its higher efficiency, lower labor and less time cost. However, there are several issues in the automatic reply method. The first as well as the most important one is that automatic reply only can be got by accurately understanding and dealing with massive information according to the question people asked. The second one is that there may be a large set of possible answers which introduces redundancies. Further, considering the difficulties of short text, high noise and high redundancy of social network messages, how to select tidy answers with high quality to fulfill users' requirement is the third problem.

Answer extraction is a key solution in this query-reply problem which generates the exact answer from the messages. Answer extraction generates candidate answers from the messages firstly and then rank them based on some scoring functions, such as frequency of the candidate, similarity between the query and candidate. Previous research has examined different methods of answer extraction such

* Corresponding author. Tel.: +86 15827135526.
E-mail addresses: pengm@whu.edu.cn (M. Peng), gbl_long@whu.edu.cn (B. Gao), zhujiahui@whu.edu.cn (J. Zhu), huangjj@whu.edu.cn (J. Huang), ymt@whu.edu.cn (M. Yuan), kevin.lifei@gmail.com (F. Li).

[1] Min Peng, PHD, major in NLP(Natural Language Process), IR(Information Retrieval).
[2] Binlong Gao, MASTER, major in NLP.
[3] Jiahui Zhu, MASTER, major in NLP.
[4] Jiajia Huang, PHD, major in NLP, IR.
[5] Mengting Yuan, PHD, major in IR.
[6] Fei Li, PHD, major in NLP.

http://dx.doi.org/10.1016/j.eswa.2015.08.056
0957-4174/© 2015 Elsevier Ltd. All rights reserved.

as Named Entity Recognition or pattern matching. However, the results are not satisfied when applying classic methods in query-reply in social network, without considering the characteristics of social network.

Most existing methods on automatic query-reply in social network pay more attention on how to generate the final reply for people's query. However, these methods always ignore the origin characteristics of social network messages such as high noise and high redundancy. As a result, the quality of reply is less than satisfactory. In this paper, we present a novel approach using information extraction and summarization method to improve the quality of the reply for users' queries in social network. The motivation behind it is as follows: First, information extraction method is used to extract high quality information from the mass information. Actually, it filters noise and removes redundancy, and can reduce the amount of messages used for computing effectively. Second, user query-oriented text summarization is studied to extract high quality information which produces a short and brief summary as the final answer. Meanwhile, in the process of summarization, noise and redundancy are reduced again, high quality sentences are organized elaborately as the answers of the questions or replies of the queries users put forward in social networks, and the quality of the summaries is ensured to attach the requirement of users.

The core idea of our method is as follows. We aim to produce query-oriented social network summarization as the reply for the query, which generates a brief and concise summary according to the keywords given by users. Compared to previous automatic query-reply methods, our work focuses on reducing noise and eliminating redundancy more effectively. We propose an information extraction method, which uses multiple features fusion and time-frequency transformation to reduce the content redundancy and extract high quality information. The method takes into account users' query, as well as multiple features of social network messages such as commented number, forwarded number, URL, content and follower number. The main process of the method is as follows. First of all, a *K*-dimensions feature matrix is constructed as an extraction basis matrix. Second, the *K*-dimensions feature is transformed into time-frequency domain to reduce computing time and improve extraction quality. And then, each feature's contribution degree of the information quality is estimated based on expectation-maximization algorithm (EM Algorithm). After the extraction, we get a set of high quality social network messages. Finally, based on the comprehensive weight of sentence's characters (e.g. content and location) and similarity between query and sentence, we score the sentences and select Top-m sentences to form the summary.

The main contributions of our work include: (1) both the information extraction and summarization method are used to give a concise and brief answer for the query automatically in social networks; (2) in the process of high quality information extraction, multiple features fusion and time-frequency transformation method are used to reduce noise and eliminate redundancy in the raw data and extract a series of high quality social network messages; (3) by scoring the sentences of the high quality social network messages we select Top-m sentences as the summary; (4) both automatic evaluation and manual evaluation are used in the evaluation of our method's results. In manual evaluation, our results are compared to human-generated summaries in terms of grammaticality, redundancy, informativeness and comprehensive quality. Results show that our method works reasonably well and outperforms other algorithms.

The outline of the paper is as follows. We discuss the related work in next section. The studied problem is defined and social network messages feature is analyzed in Section 3. In Section 4 the information extraction method is described. The summarization of social network messages is introduced in Section 5. Experimental results and evaluation are shown in Section 6. Finally, we conclude our work.

2. Related work

Query-reply or question answering (QA) is studied as a fundamental problem by the machine learning and the artificial intelligence communities. Previous studies on query-reply have discussed utilizing various methods for answer extraction, including patterns, neural network, passage graph, etc. Soubbotin (2001) used hand-crafted patterns to extract candidates from the documents for pre-defined question type and Ravichandran and Hovy (2002) enriched the method by adding semantic types to the question terms. The modified method given by Ravichandran used the automatically learned patterns as features to model the correctness of the extracted answers. In contrast to the aforementioned approaches, Iyyer (2014) introduced a dependency tree recursive neural network recursive neural network model (QANTA) which placed the burden of learning answer types and patterns on itself by learning word and phrase-level representations that combine across sentences to reason about entities to give the best answer. In the study of Sun, Duan, and Duan (2013), answer extraction was performed on a Passage Graph which was built by linking words with the same stem upon all the passages retrieved for the same question, for the purpose of increasing the possibility for correctly answering the question. As the generation of paraphrases is a useful means to improve the search for finding best answers in community question answering (cQA), Figueroa and Neumann (2013) used query logs from *Yahoo! Search* and *Yahoo! Answer* for automatically extracting a corpus of paraphrases of queries and questions, and then learned to rank models to recognize effective search queries so as to fetch answers from cQA services. Compared to the methods above, in addition, our method put forward generating a concise and brief summary automatically as the answer for the query in social networks for the sake of its high quality.

Zhou (2014) presented an enhanced hybrid approach to OWL query answering that combines an RDF trip-store with an OWL reasoner in order to provide scalable pay-as-you-go performance. They exploited summarization techniques to prune candidate answers in order to "shrink" the data in a knowledge base. However, they only used summarization method to compress the data. Different from their method, in this paper, we aim to extract useful information in the first step and summarize the candidate answers, which give user a non-original answer but a brief and accurate summary as the final answer.

In terms of information extraction, traditional methods mainly consider content feature of information and they extract topic relevant information mainly based on LDA model, TF-IDF model or other topic models (Daubechies, 1992; Sharifi, Huttion, & Kalita, 2010a; Xia, He, Tian, Chen, & Lin, 2011). However, most of these efforts didn't take into consideration the social media features completely. They only referred to one or two social media features such as poster authority feature. Other researchers pay attention to other variety of social media features such as poster influence, tags, URL and time to improve extraction precision. For example, Harabagiu and Hickl (2011) believed that user behavior towards individual tweets could be used to identify relevant content for social network information extraction, so they focused on three kinds of chains: retweet chains, responds chains and quote chains and accessed the relevance of content expressed by any group of users linked by the chains. Although this method have considered several aspects and performed well, they haven't analyzed some of the features which were really important for information extraction. Along this line of work, we also construct a set of features for sentence salience ranking. Yet our method tries to employ and model the relationships among more relevant and necessary features globally to produce summary with diversity and sufficient coverage.

Many research studies also focus on social network summarization in the last few years: the task of selecting a list of meaningful social network posts that are most representative for some topic.