Neurocomputing 285 (2018) 117–124

Contents lists available at ScienceDirect

Neurocomputing

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

Question retrieval for community-based question answering via heterogeneous social influential network

Zheqian Chen*, Chi Zhang, Zhou Zhao, Chengwei Yao, Deng Cai

State Key Lab of CAD & CG, Zhejiang University, Hangzhou, Zhejiang 310058, PR China

ARTICLE INFO

Article history:
Received 12 July 2017
Revised 29 December 2017
Accepted 15 January 2018

Communicated by Guan Ziyu

Keywords:
CQA
Question retrieval
Deep learning
Social network

ABSTRACT

Community-based question answering platforms have attracted substantial users to share knowledge and learn from each other. As the rapid enlargement of community-based question answering (CQA) platforms, quantities of overlapped questions emerge, which makes users confounded to select a proper reference. It is urgent for us to take effective automated algorithms to reuse historical questions with corresponding answers. In this paper, we focus on the problem with question retrieval, which aims to match historical questions that are relevant or semantically equivalent to resolve one's query directly. The challenges in this task are the lexical gaps between questions for the word ambiguity and word mismatch problem. Furthermore, limited words in queried sentences cause sparsity of word features. To alleviate these challenges, we propose a novel framework named *HSIN* which encodes not only the question contents but also the asker's social interactions to enhance the question embedding performance. More specifically, we apply random walk based learning method with recurrent neural network to match the similarities between asker's question and historical questions proposed by other users. Extensive experiments on a large-scale dataset from a real world CQA site Quora show that employing the heterogeneous social network information outperforms the other state-of-the-art solutions in this task.

1. Introduction

Community-based question answering (CQA) services enable users to put forward their puzzles and share knowledge with each other. Over the past years, CQA services like Yahoo! Answers, Baidu Knows, Wiki Answers, Zhihu and Quora have accumulated substantial question-answer pairs [1]. However, large quantities of proposed questions are highly overlapped and redundant, which weakens users query efficiency [2]. To effectively automate select the proper references from the large-scale pre-queried questions with corresponding answers, researchers have devoted into question retrieval, question answering, expert finding and natural language processing field for many years.

In this paper, we focus on the domain of question retrieval. The critical problem of question retrieval is to help users to retrieve historical questions which precisely match their questions semantically equivalent or relevant [3]. Users can refer to the good matches before choosing whether to raise a new question. The functionality brings users much convenience and reduces the repetition rate for CQA platforms. Hence, it is of great value for CQA services to offer relevant results efficiently and precisely. Many studies have been done on this task. However, challenges still remain due to the lexical gaps between questions caused by word ambiguity and word mismatch problem [4]. For example, in Quora site there exit two questions “What are some good introductory materials on machine learning?” and “How can I start learning machine learning?”. From the views of our human readers, these two questions are semantically relevant and exactly express the same meaning. While for the main stream models applied in question retrieval, these two questions share few common words so they may cause the mismatching problem. Even for the same word may cause ambiguity, for instance when we mention the word ‘apple’, we cannot easily tell whether it is about the apple company or the apple fruit [5] unless we classify through context information. Another challenge in question retrieval is the feature sparsity issue [6]. As question titles usually have short length with varieties of irregular noise, it is hard to extract exactly modeling topic from using the full information of questions.

Most of the existing works consider the question retrieval task as a supervised learning method, which utilizes both the question textual content and its belonging category to train an evaluating model [7–9]. Researchers in question retrieval field mainly exploit the language model to learn the semantic representation of

* Corresponding author.
E-mail addresses: zheqianchen@gmail.com (Z. Chen), wellyzhangc@zju.edu.cn (C. Zhang), zhaozhou@zju.edu.cn (Z. Zhao), yaochw@zju.edu.cn (C. Yao), dengcai@gmail.com (D. Cai).

https://doi.org/10.1016/j.neucom.2018.01.034
0925-2312/© 2018 Elsevier B.V. All rights reserved.

question contents. Although the existing question retrieval methods have achieved excellent performance, they do not fully tackle the word sparsity bottleneck and utilize the questions side information such as asker's background, which is critical for question understanding. Moreover, since askers have their own social network and their interests may be similar with their friends, it is reasonable to assume one scenario: users may post questions resembled with their friends. It is a very common phenomenon among classmates and colleagues. Thus, how to leverage these available social information is of significance for the question retrieval task.

Apart from the valuable social information, the textual contents of questions are necessary for question retrieval tasks. Recent works on question retrieval in CQA data employ different retrieval models to learn semantic representations, including the language model [10], the translation model [11] and learning-to-rank model [9]. Empirically, these previous works consistently show the feasibility in retrieval performance. However, traditional hand-crafted features like bag-of-words have inevitable issues that cannot well-embed the word sequence of questions. Inspired by the flourish of deep learning application in natural language processing [12], various embedding methods are proposed for learning the semantics of similar words and encode the word sequence into low-dimensional continuous embedding space. Since the question contents are always sequential data with variant length, recurrent neural network [13] is an ideal choice to learn the semantic representation.

In this paper, we put forward a novel framework named HSIN (Heterogeneous Social Influential Network). Specifically, we exploit a random walk method to explore the valuable side information from heterogeneous social network and question category information. Besides, we model the question textual content with recurrent neural network. We then concatenate the question textual content with user embedding to represent the question and rank similarities with historical questions. In our proposed HSIN framework, the questions textual content, their related categories information, askers social information are simultaneously learned so we can utilize the rich interactions between CQA data and users data. When a new question is queried, HSIN can rank the historical proposed similar questions so that users can refer to the recommended questions along with corresponding answers without having to wait his own question to be answered.

It is worthwhile to highlight several contributions of our work here:

- We introduce a novel framework named HSIN to integrate question textual content with asker social network information. We utilize a random deep walk method with recurrent neural network to learn the semantic representation of questions and users simultaneously.
- Unlike previous studies, our proposed framework which leverages the semantic representation of questions and the rich heterogeneous social network information in question retrieval field. The framework can be extended to other information retrieval field for it is scalable for heterogeneous network learning.
- Our proposed framework outperforms the state-of-the-art models that utilized only question textual information. The performance improved significantly in question retrieval ranking, which demonstrate the potential of our concept of integrating the rich social network side information.

The remainder of this paper is organized as follows. In Section 2, we present a brief view of current related work about question retrieval. In Section 3, we formulate the question retrieval problem and introduce our proposed heterogeneous network integration learning method. In Section 4, we describe the experimental settings and report a variety of results to verify the superiority of our model. Finally, we conclude the paper in Section 5.

2. Related work

The existing methods for question retrieval can be basically categorized as categories-model based approaches, translation-model based approaches, topic-modeling based approaches and neural network based approaches.

The first approach is the most widely considered in exploring question retrieval problems. It considers the metadata of questions by taking question categories and labels into consideration. Cao et al. [7,14,15] embodied three language models to exploit question categories smoothing for estimating questions similarities under the same category. Zhou et al. [4,10,16] proposed several methods in employing category side-information. In [16], they leveraged user chosen category and filter irrelevant questions under leaf categories. In [10], they developed group non-negative matrix factorization with learning the category-specific topics for each category as well as shared topics across all categories. Zhou et al. [4] also employed fisher kernel to aggregate word embedding vectors from variable size into fixed-length, thus learnt a continuous word embedding model.

The second approach, translation-model based method learns the pair relevance of question-answer data to bridging Lexical gaps between queries and questions, or questions and answers. Jeon et al. [3] discussed a method that refers to the similarities between answers to estimate question semantically similar probabilities. Xue et al. [17] combined the question part with a query likelihood approach by incorporating word-to-word translation probabilities. Lee et al. [18] investigated empirical methods to eliminate non-topical or unimportant words in order to construct compact translation models for retrieval purposes. Apart from word-level translation method, Zhou et al. [11] learnt a phrase-based translation model which aims to capture question contextual information rather than word-based in isolation.

The third approach topic modeling based approaches also arose many attentions for we can compare the latent similarity of questions without being constrained by the queried sentences forms. Duan et al. [19] identified the question topic and focus into a consisting data structure. Zhang et al. [20] assumed that questions and answers share some common latent topics and through this way the model can match questions on a topic level. Ji's et al. [21] assumption is quite similar to Zhang et al. [20].

The fourth approaches leverage neural network to model questions embeddings. As the flourish of deep learning especially in natural language processing, researchers began to incorporate the neural network into learning to rank frameworks. Zhou et al. [22] learnt the semantic representation of queries and answers by using a neural network architecture. Although the mainstream models of neural network are mainly applied in question answering not in question retrieval, the theories are the same. Qiu and Huang [23] encoded questions and answers in semantic space and model their interactions in a convolutional neural tensor network architecture. The model is a general architecture with no need for lexical or syntactic analysis. Shen et al. [24] utilized a similarity matrix which contains both lexical and sequential information to effectively model the complicated matching relations between questions and answers.

In question retrieval field, very few approaches consider the heterogeneous social network to dig more information. For example, classmates or colleagues concern the same professional field so they may care about the similar questions. Zhao et al. [25] implemented a graph-regularized matrix completion algorithm by integrating the user model to improve expert finding performance. The cross-domain social information integration is also considered