Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)

Community-Based Question Answering via Contextual Ranking Metric Network Learning

Hanqing Lu, Ming Kong
{lhq110, zjukongming}@zju.edu.cn
College of Computer Science, Zhejiang University
No.38 Zheda Road, Hangzhou, Zhejiang, China, 310027

Abstract

The exponential growth of information on Community-based Question Answering (CQA) sites has raised the challenges for the accurate matching of high-quality answers to the given questions. Many existing approaches learn the matching model mainly based on the semantic similarity between questions and answers, which can not effectively handle the ambiguity problem of questions and the sparsity problem of CQA data. In this paper, we propose to solve these two problems by exploiting users' social contexts. Specifically, we propose a novel framework for CQA task by exploiting both the question-answer content in CQA site and users' social contexts. The experiment on real-world dataset shows the effectiveness of our method.

Introduction

The benefits of CQA have been well-recognized today. The problem of automatic question answering in CQA sites is to rank the relevant answers to the given questions posted by the users, which has attracted a lot of attention recently. Most of the existing works consider the problem of question answering as the short-text matching task. Although existing question answering methods have achieved promising performance, most of them still suffer from the following two problems: *1. The sparsity of the CQA data*: The links between questions and answers are sparse. *2. The ambiguity of questions*: Questions can be ambiguous and may be ill-received by users. However, existing methods, which only utilize the content of questions and answers, can hardly solve the above two problems.

Fortunately, with the prevalence of online social media, many social media platforms host much data that can be used to improve the performance of CQA, such as users' social connection and their posted social content (e.g., tweets and retweets) in various online social networks (e.g., Twitter). By using the available information, we are able to alleviate the above two problems of CQA. In this paper, we introduce the problem of question answering in CQA site from the viewpoint of contextual ranking metric network embedding. Specifically, we propose the heterogeneous CQA network that integrates both question-answer content in CQA site and users' social contexts into a unified Contextual Ranking Metric Network Learning framework, named as CRMNL. We then develop a random-walk based learning method with deep recurrent neural networks to learn the representations of questions, answers, and users, such that the ranking metric is implicitly embedded in the representations. When a certain user asks a question, CRMNL can rank the answers for this user based on the trained ranking metric embedding.

The Framework

We denote the proposed heterogeneous CQA network by $G = (V, E)$ where the set of nodes V is composed of questions X, answers Y and users' social contexts Z, and the set of edges consists of relative quality rank T and social relations S. We illustrate a simple example of ranking metric heterogeneous CQA network modeling in Figure 1(a). The question $\mathbf{q}_1$ asked by user $\mathbf{u}_1$ has a high-quality answer $\mathbf{a}_1$ (i.e., marked with $+$ on the answering relation) and a low-quality answer $\mathbf{a}_2$ (i.e., marked with $-$ on the answering relation). The matching quality is voted through thumbups/downs, which indicates the community's long term review. We also illustrate the following relation between users $\mathbf{u}_1$ and $\mathbf{u}_2$ in Figure 1(a).

Inspired by DeepWalk (Perozzi, Al-Rfou, and Skiena 2014), we sampled the paths from the heterogeneous CQA network by using random-walk method. We considered them as the context windows for the vertex embedding in networks. For example, the context of vertex v_i with the c-th sampled window is $W_c = \{v_{i-\frac{w}{2}}, \ldots, v_{i-1}, v_{i+1}, v_{i+\frac{w}{2}}\}$, where the length of sampled path is w.

To leverage the supervised information, we integrate the random-walk method in DeepWalk with the recurrent neural networks based learning into a unified CQA network learning framework. Specifically, we train the LSTM (Hochreiter and Schmidhuber 1997) networks for questions, answers and users' textual content, which are named Q-LSTM, A-LSTM, and U-LSTM, respectively. As illustrated in Figure 1(c), to handle the question ambiguity problem, we introduce the concatenation layer for question representation under users' social context, which concatenates the representation of a given question and the representation of the user's textual content into one embedding vector.

Given the context windows W and embedding vectors for all vertices in W, for each vertex v_i in the context window W, its loss function is given by:

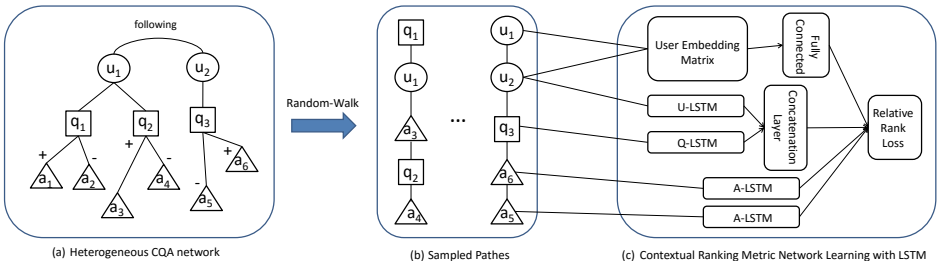

Figure 1: Example of heterogeneous CQA-network

$$l(v_i) = \begin{cases} \sum_{a^+,a^- \in W} \max(0, m + f_{\mathbf{u}}(\mathbf{v}_i, \mathbf{a}^-) - f_{\mathbf{u}}(\mathbf{v}_i, \mathbf{a}^+)) \\ , \; if \; v_i \in Q \\ \sum_{u \in W} \|u - v_i\|^2 \\ , \; if \; v_i \in U \end{cases} \quad (1)$$

where the superscript a^+ denotes the high-quality matching answer (with higher votes) and a^- denotes the low-quality matching answer (with lower votes) for question v_i. We denote that function $f_{\mathbf{u}}(\mathbf{q}, \mathbf{a})$ quantifies the matching quality between the answer $\mathbf{a}$ and the question $\mathbf{q}$ asked by the user $\mathbf{u}$. We denote the hyper-parameter m $(0 < m < 1)$ controls the margin in the loss function. Q and U are the sets of questions and users, respectively.

Therefore, the objective function is given by:

$$\min_{\Theta} L(\Theta) = \sum_{W} \sum_{v_i} l(v_i) + \lambda \|\Theta\|^2 \quad (2)$$

where Θ denote all the model parameters, and λ is the trade-off parameter between the training loss and regularization. To optimization the objective function, we employ the stochastic gradient descent (SGD).

Experiments

We evaluate the performance of our method by using the Quora dataset (Zhao et al. 2015). The dataset contains 444,138 questions, 95,915 users, 887,771 answers from Quora, and users' following relationship in Twitter. We also crawl the Quora user's posted tweets on Twitter since the dataset provides the Twitter Ids for Quora users. The quality of users on answering the question is indicated through thumbs-up/down voted by the community. We evaluate the performance of different methods based on Accuracy with different ratio of training data from 60%, 70% to 80%. We compare our method with three baseline methods for the problem of CQA as follows:

BOW method uses the bag-of-words representation of both questions and answers for computing the relevant score.

Doc2Vec method (Le and Mikolov 2014) uses the distributed bag-of-words representation that encodes questions and answers into a low-dimensional feature space.

DeepWalk method (Perozzi, Al-Rfou, and Skiena 2014) learns the embedding of both questions and answers based on the network structure.

Table 1 shows the evaluation results of all the question answering methods on Accuracy. The hyperparameters and parameters which achieve the best performance on the validation set are chosen to conduct the testing evaluation. We then report the average value of all the methods. The experiments demonstrate that our proposed CRMNL method achieves the best performance in all the cases. This result illustrates that the proposed framework that employs both deep representation of questions and answers, as well as users' context can further improve the performance of question answering.

Table 1: Experimental results on Accuracy.

Ratio	BOW	Doc2Vec	DeepWalk	CRMNL
60%	0.3463	0.2991	0.3034	**0.4033**
70%	0.3911	0.3470	0.3318	**0.4781**
80%	0.4347	0.4187	0.3759	**0.5423**

Conclusion

In this paper, we introduced the problem of CQA from the viewpoint of contextual ranking metric network learning. We propose the heterogeneous CQA network that exploits both question-answer content in CQA site and users' social contexts for question answering. We then develop a novel contextual ranking metric network learning framework based on the heterogeneous CQA network.

Acknowledgement

This work was supported by National Natural Science Foundation of China under Grant 61602405, and the Fundamental Research Funds for the Central Universities 2016QNA5015.

References

Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. *Neural Computation* 9(8):1735–1780.

Le, Q. V., and Mikolov, T. 2014. Distributed representations of sentences and documents. *arXiv preprint arXiv:1405.4053*.

Perozzi, B.; Al-Rfou, R.; and Skiena, S. 2014. Deepwalk: Online learning of social representations. In *SIGKDD*, 701–710. ACM.

Zhao, Z.; Zhang, L.; He, X.; and Ng, W. 2015. Expert finding for question answering via graph regularized matrix completion. *IEEE Trans. Knowl. Data Eng.* 27(4):993–1004.