Construction and Completion of Document-Level Multimodal Question and Answer Knowledge Graph

Xiaoyi Zhang

School of Economics and Management, China University of Petroleum-Beijing, China

13669001077@163.com

Abstract. To make full use of the documents in question and answer (Q&A) community with text and image information included, improving the ability of information retrieval and semantic understanding, this paper focuses on the construction and completion of a multimodal Q&A knowledge graph. Firstly, we propose a document-level multimodal question and answer knowledge graph (DMQAKG), using topic, question, and answer documents as nodes, and building document relations. Furthermore, we also propose a multimodal Q&A knowledge graph completion method (MQAKGC) for DMQAKG based on multimodal feature extraction and fusion and multimodal knowledge graph link prediction. We use graph convolutional network (GCN) to capture the constructional features and long-short term memory (LSTM) to learn the chronological dependency between the entities for further completion of the missing relations. Experimental results show the superior performance of the proposed knowledge graph completion method in different Q&A subset scales.

Keywords: Multimodal Question and Answer Knowledge Graph; Knowledge Graph Completion; Virtual Q&A Community.

1 Introduction

In virtual question and answer (Q&A) communities, users can not only post text information but also share image information, which enriches the representation of traditional text features. However, the increasing internet resources lead to information overload and users are difficult to obtain information in need efficiently. Applying a knowledge graph to virtual Q&A community can help users get answers more conveniently. Multimodal knowledge graphs add image data based on traditional knowledge graphs, providing auxiliary semantic information[1]. In this context, knowledge graph completion seems to be particularly important[2,3], thus updating and enlarging knowledge graph by new entity recognition and relation extraction. In virtual Q&A community, users may raise several new questions, which may involve relations that are not covered in the knowledge graph. Predicting the relation between two entities by link prediction can enhance the coverage and accuracy of the knowledge graph, and make better use of information within[4].

K. Zhang et al. (eds.), *Proceedings of the 5th International Conference on Economic Management and Big Data Application (ICEMBDA 2024)*, Advances in Economics, Business and Management Research 313,
https://doi.org/10.2991/978-94-6463-638-3_46

This paper proposes to construct the document-level multimodal question and answer knowledge graph (DMQAKG) combining the characteristics of virtual Q&A community. And a multimodal Q&A knowledge graph completion method (MQAKGC) for DMQAKG is further proposed based on multimodal feature extraction and fusion and multimodal knowledge graph link prediction. In MQAKGC, multimodal feature fusion is conducted based on multimodal feature similarity, complementarity, and dominant dependency. The modal achieves the representation of nodes and edges through graph convolutional network (GCN), extracts potential relation features among triples through long-short term memory (LSTM) network, and constructs new links between documents. This infers new relations between entities and purposefully selects relevant, high quality, and comprehensive Q&A lists for users from a huge number of Q&A.

2 Related Works

Users usually post multimodal unstructured data at the same time in the virtual Q&A community. Fusing data information from different modalities provides richer and more comprehensive information for the construction of a knowledge graph. Measuring the distance between different modalities generates a public subspace[5]. To capture image information effectively, Benitez-Quiroz et al.[6] combined local and global features and used a novel information transmission algorithm between classes for optimization.

Completion of knowledge graph refers to the continuous modification, updating and expansion of knowledge graph with new entity recognition and relation extraction[7]. Researchers have explored the introduction of deep learning methods into knowledge graph completion, with research on link prediction based on graph neural networks becoming a hot spot. Link prediction enhances the accuracy and efficiency of predicting links between entities by embedding both document entity features and relationship features into a lower dimensional space. Nguyen et al.[8] propose a personalized topic recommendation system based on convolutional neural networks, which utilizes convolution and maximum pooling layers to gain visual features in images.

3 Methodology

3.1 Construction of DMQAKG

During the process of constructing DMQAKG, entity recognition and relation extraction are two key steps. Entity recognition uses question, topic and answer documents in virtual Q&A community as basic entities. Establishing corresponding relations is needed from the answer to the question, from the answer to the topic, and from the topic to the question. Meanwhile, relation extraction calculates the similarity between the contents of the document entities. The structure of DMQAKG is shown as Fig. 1.

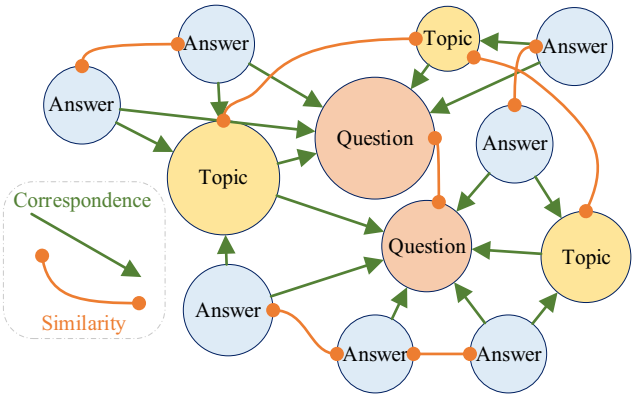

Fig. 1. The Structure of DMQAKG

3.2 Completion of DMQAKG

Based on the constructed DMQAKG, this paper proposes a completion method (MQAKGC) for DMQAKG based on multimodal feature extraction and fusion and multimodal knowledge graph link prediction.

Multimodal Feature Extraction and Fusion. To make full use of the text and image information of documents in DMQAKG, multimodal feature extraction and fusion capture potential semantic information of different data sources, facilitating a better understanding of relations between nodes. Text feature extraction first needs to extract features from vocabulary in the documents and vectorize the representations, to convert the original natural language text data into feature representation that can be recognized by computer. Image feature extraction needs to utilize convolutional neural networks to extract abstract local features of image documents.

In multimodal fusion framework, three fusion metrics are used: feature similarity, feature complementarity and feature dominant dependency. Multimodal feature similarity uses cosine distance to measure similarity between multimodal feature vectors, as shown in Eq. (1). Multimodal feature complementarity measures the difference in semantic information that may be included between multimodal features, as shown in Eq. (2). Multimodal feature dominant dependency measures which modal has stronger control over the final semantic representation in the multimodal fusion process, as shown in Eq. (3).

$$SIM(Xt, Xi) = \frac{Xt \cdot Xi}{\|Xt\| * \|Xi\|} \tag{1}$$

$$COM(X, Xt, Xi) = max\big(dis(X, Xt), dis(X, Xi)\big) \tag{2}$$

$$Ut = St + Si + Ct, Ui = Ct + Ci + Si \tag{3}$$

Where Xt and Xi represent the feature vectors of text and image. X represents the intersection of feature Xt and Xi in the feature subspace. Ut and Ui represent the dominant dependency of text and image features. St and Si denote the shared features

of text and image, while Ct and Ci represent the complementary features of text and images.

Multimodal Knowledge Graph Link Prediction. Encoding the entities in the triples is needed to generate node embedding. An adjacency matrix of graphic data is generated according to relation information in the triples. To prevent overfitting, negative samples need to be generated for each entity and relation in the training set, which cannot match the existing links. Negative sample generation plays an important role in the completion of knowledge graph based on link prediction, increases the generalization capability of the model and reduces overfitting.

The encoder based on GCN can process knowledge graph structures with convolutional layers and generate node embeddings to extract local context information from the knowledge graph [9]. It integrates node information in virtual Q&A community into average embedding vectors of the whole graph, which can capture the relations among nodes. During this process, information from each node needs to be converted into variable level information, a weighted average for each node is computed at each time step, and concatenate these weighted averages into a variable vector. The integration calculation in the GCN encoder is shown in Eq. (4).

$$H = \sum_{i=1}^{n} \alpha_i h_i \tag{4}$$

Where α_i represents the weight coefficient of node i. h_i denotes the embedding vector of node i. And n represents the total number of nodes in the graph. Pooling operation can capture the global structure of the graph better and increase the prediction performance of the model.

The decoder based on LSTM can predict the next node embedding vector and the final generated output sequence according to a node embedding vector sequence generated by the encoder[10]. This method can help to understand and capture the dynamic variation and relevance between nodes better, facilitating more accurate and reliable link prediction results. In each time step, the node embedding vectors and the LSTM output vector from the previous time step are fed into the LSTM decoder, while cell state vectors are generated and updated, as shown in Eqs. (5)-(7).

$$\overline{c_t} = \tanh(A_t W_c + GCN_o K(A_{t-1}, h_{t-1}) + b_c) \tag{5}$$

$$i_t = \sigma(A_t W_i + GCN_c K(A_{t-1}, h_{t-1}) + b_i) \tag{6}$$

$$c_t = f_t \odot GCN_c K(c_{t-1}) + i_t \cdot \overline{c_t} \tag{7}$$

Where $\overline{c_t}$ represents the current candidate cell state vector. c_t and c_{t-1}denote the cell state vectors at the previous and current time steps. A_t and A_{t-1} refer to the node feature matrices at the current and previous time steps. $K(A_{t-1}, h_{t-1})$ represents the combination of the node feature and hidden state from the previous time step. $W_{c,i}$ and $b_{c,i}$ are the weight and bias matrices of the input gate.

Whether there is a relation between two entities and the relation intensity can be predicted through the link prediction method above.

4 Experiment

The dataset is constructed from one of the most popular and widely used virtual Q&A communities in China, Zhihu. It consists of 7,115 valid data, including 40 questions, 68 topics, and 7115 answers. The proposed knowledge graph DMQAKG consists of 300 entities and 21862 relations, including 40 questions, 68 topics, and 192 answers.

To evaluate the MQAKGC model, we selected nine knowledge graph completion models for comparison: Logistic Classifier (A), Decision Trees (B), Random Forests (C), K Nearest Neighbors (D), Gaussian Naive Bayes (E), Support Vector Classifiers (F), Multilayer Perceptron (G), Adaptive Boosting (H) and Gradient Boosting (I). This paper sets 15%, 10% and 5% scales of Q&A subsets for link prediction. The knowledge graph completion performance of these models was compared and analyzed using the four metrics, including Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Mean Square Error (MSE), and Root Mean Square Error (RMSE).

Table 1. Link Forecast Model Performance

	15%				10%				5%			
Model	MAE	MAPE	MSE	RMSE	MAE	MAPE	MSE	RMSE	MAE	MAPE	MSE	RMSE
MQAKGC	**0.06**	**0.48**	**0.04**	**0.21**	**0.01**	**0.30**	**0.03**	**0.19**	**0.07**	**0.34**	**0.02**	**0.15**
A	0.46	1.41	0.27	0.52	0.50	1.66	0.30	0.55	0.48	1.42	0.28	0.53
B	0.48	1.45	0.27	0.52	0.49	1.42	0.28	0.53	0.54	1.80	0.34	0.58
C	0.52	1.90	0.32	0.56	0.52	1.96	0.32	0.56	0.50	1.47	0.29	0.54
D	0.39	0.98	0.20	0.45	0.41	0.97	0.21	0.46	0.42	1.04	0.21	0.46
E	0.43	1.21	0.23	0.48	0.41	1.24	0.22	0.47	0.36	1.31	0.20	0.44
F	0.49	1.48	0.28	0.52	0.48	1.62	0.28	0.53	0.39	1.35	0.22	0.47
G	0.49	1.48	0.28	0.52	0.48	1.62	0.28	0.53	0.39	1.35	0.22	0.47
H	0.47	1.61	0.27	0.52	0.52	2.05	0.32	0.57	0.55	2.23	0.34	0.58
I	0.47	1.65	0.28	0.53	0.54	1.95	0.33	0.57	0.49	1.39	0.27	0.52

Among the models in the comparison experiment above, the proposed MQAKGC has the best performance, as presented in Table 1. MQAKGC makes full use of key information in the document resources of DMQAKG from the perspective of multi-modal information fusion. The structure of MQAKGC is more complex and requires more parameters. In summary, MQAKGC has high accuracy, low relative error and small overall error in knowledge graph completion tasks.

To evaluate the effect of Q&A subset scales on the error of link prediction models, this paper conducts a performance comparison under three different Q&A subset sizes to evaluate the effectiveness of MQAKGC in knowledge graph completion tasks. When the scale of the dataset is reduced from 15% to 10% and 5%, most of the comparison models show an increase in error metrics, which indicates that they are more sensitive to a reduction in the number of data samples. Meanwhile, MQAKGC performs relatively stable, which shows its strong generalization ability and adaptability.

5 Conclusions

Users in virtual Q&A communities often post Q&As including both text and image modal data. To fuse multimodal information and integrate Q&A document data, this paper focuses on the construction and completion of a multimodal question and answer knowledge graph. To integrate multimodal document information containing images and text, we construct the multimodal Q&A knowledge graph DMQAKG. To address the issue of content gaps in the knowledge graph, this paper proposes a knowledge graph completion method MQAKGC based on multimodal extraction and fusion and multimodal knowledge graph link prediction. In the MQAKGC, firstly, multimodal feature similarity, multimodal feature complementarity, and multimodal feature dominant dependency are used for feature fusion. Then, GCN is used to capture the features of knowledge graph nodes and edges, while LSTM is used to discover the potential relation features between triples. It discovers new relations between entities continuously, integrating these new entities and relations into the knowledge graph, and adapts the constant altering and updating of the knowledge graph effectively. The experiment results show that the proposed model has the best performance. In future work, more datasets from other virtual Q&A communities and more link prediction algorithms for extended comparison experiments will be utilized to further validate the practicality, adaptability, and generalization ability of our proposed method.

Acknowledgment

This research was funded by the National Key Research and Development Program of China under Grants No. 2021YFF0600401.

References

1. S. Ji, S. Pan, E. Cambria, P. Marttinen, and P. S. Yu, “A Survey on Knowledge Graphs: Representation, Acquisition and Applications,” *IEEE Trans. Neural Netw. Learning Syst.*, vol. 33, no. 2, pp. 494–514, Feb. 2022, doi: 10.1109/TNNLS.2021.3070843.
2. C.-H. Lee, D. Kang, and H. J. Song, “Fast knowledge graph completion using graphics processing units,” *Journal of Parallel and Distributed Computing*, vol. 190, p. 104885, Aug. 2024, doi: 10.1016/j.jpdc.2024.104885.
3. J. Leblay and M. W. Chekol, “Deriving Validity Time in Knowledge Graph,” *The Web Conference 2018 - Companion of the World Wide Web Conference, WWW 2018*, pp. 1771–1776, 2018, doi: 10.1145/3184558.3191639.
4. J. Chen, X. Wang, and X. Xu. “GC-LSTM: graph convolution embedded LSTM for dynamic network link prediction,” *Applied Intelligence*, vol. 52, pp. 7513–7528, Sep. 2022, https://doi.org/10.1007/s10489-021-02518-9
5. N. N. Daud, S. H. Ab Hamid, M. Saadoon, F. Sahran, and N. B. Anuar, “Applications of link prediction in social networks: A review,” *Journal of Network and Computer Applications*, vol. 166, p. 102716, Sep. 2020, doi: 10.1016/j.jnca.2020.102716.

6. S. Tao, R. Qiu, Y. Ping, and H. Ma, "Multi-modal Knowledge-aware Reinforcement Learning Network for Explainable Recommendation," *Knowledge-Based Systems*, vol. 227, p. 107217, Sep. 2021, doi: 10.1016/j.knosys.2021.107217.
7. W. Zaremba, I. Sutskever, and O. Vinyals, "Recurrent Neural Network Regularization," Feb. 19, 2015, *arXiv*: arXiv:1409.2329. doi: 10.48550/arXiv.1409.2329.
8. C. Yu, X. Zhao, L. An, and X. Lin, "Similarity-based link prediction in social networks: A path and node combined approach," *Journal of Information Science*, vol. 43, no. 5, pp. 683–695, Oct. 2017, doi: 10.1177/0165551516664039.
9. D. Peng and Y. Zhang, "MA-GCN: A Memory Augmented Graph Convolutional Network for traffic prediction," *Engineering Applications of Artificial Intelligence*, vol. 121, p. 106046, May 2023, doi: 10.1016/j.engappai.2023.106046.
10. M. A. Kosan, H. Karacan, and B. A. Urgen, "Predicting personality traits with semantic structures and LSTM-based neural networks," *Alexandria Engineering Journal*, vol. 61, no. 10, pp. 8007–8025, Oct. 2022, doi: 10.1016/j.aej.2022.01.050.