Transactions on Computational and Scientific Methods | Vol. 4, No. 12, 2024
ISSN: 2998-8780
https://pspress.org/index.php/tcsm
Pinnacle Science Press

Enhancing Recommendation Systems through a Multi-Modal Knowledge Graph Neural Network Model

Zayne Alcott
Illinois Institute of Technology, Chicago, USA
zayne58@iit.edu

Abstract:

With the rapid development of artificial intelligence, recommendation systems play a crucial role across various domains such as entertainment, e-commerce, and social networking. Collaborative filtering has traditionally been the cornerstone of recommendation algorithms but faces limitations due to data sparsity and the cold start problem. In recent years, the introduction of knowledge graphs has enhanced recommendation systems by providing structured auxiliary information; however, these models typically rely solely on text-based data, overlooking other valuable forms of information such as images, audio, and video. This study addresses this gap by proposing a recommendation model based on a multi-modal knowledge graph neural network (MGNN), integrating text, visual, and audio data to create a more comprehensive and dimensional knowledge representation. Focusing on the domain of movie recommendations, we construct a multi-modal knowledge graph and employ the MGNN model to fuse features from diverse data types. This enables the extraction and aggregation of multi-modal attributes, thereby enhancing recommendation accuracy and system performance. Experimental results demonstrate that the multi-modal knowledge graph approach substantially outperforms traditional recommendation systems. This research contributes to the fields of multi-modal data integration and knowledge graph-enhanced recommendation systems, and lays groundwork for further advancements in multi-modal recommendation methodologies.

Keywords:

Multi-modality, knowledge graph, recommendation system, click rate.

1. Introduction

With the development of artificial intelligence, online recommendation systems are widely used in various fields of life such as entertainment, social interaction, food, etc. Therefore, the update of recommendation algorithms is also emerging. Many giant companies, e-commerce and entertainment companies are vigorously Develop recommendation algorithms, try to use recommendation algorithms to predict user preferences, increase click-through rates and conversion rates, and empower enterprises. Among the recommendation strategies, collaborative filtering (CF), as the most classic algorithm strategy, is mainly based on the historical interaction behavior of users with similar characteristics to recommend items for users [1].[2], attribute information [3], image information [4] and context information [5], etc.

In recent years, in order to recommend more comprehensive information and higher accuracy of recommendations, knowledge structures such as knowledge graphs have been widely used in recommendation systems [6,7]. However, the existing recommendation algorithms based on knowledge graphs use only the text-only graph structure to a large extent, ignoring the multi-modal information describing the item information, such as the text, audio, image and other information

describing the object. For example, in the field of video recommendation, the original knowledge graph only contains text information such as movie name, actor, movie genre, director, etc. However, there will be a lot of similar movies based on these characteristics, especially in the series of movies, in order to be more precise. To recommend movies to users, we need to add more types of auxiliary information to achieve a multi-dimensional description of movies. Therefore, multi-modal graphs have become an inevitable choice. Through multi-modal knowledge graphs, not only the connections between entity concepts and attributes can be obtained, but also the connections between different modal data can be obtained. So as to get a higher-dimensional description of the object to be recommended.

Traditional knowledge graph technology has been widely used to process structured data (using ontology + D2R technology) and text data (using text information extraction technology), but there is also a type of unstructured data, namely visual data, audio data, etc., which is relatively The degree of attention is low, and there is a lack of effective technical means to extract structured knowledge from these data. In recent years, although some multi-modal technologies have been proposed, these technologies are mainly to improve the effects of image classification, image generation, and image question and answer, and cannot well support the construction of multi-modal knowledge graphs.

In 2007, Auer S proposed DBpedia[8] as the central field of semantic web research in the past decade. Its rich semantic information is the link endpoint of the current multi-modal knowledge graph, and the complete ontology structure is essential for constructing multi-modal knowledge. Atlas provides great convenience. In 2014, Wikidata [9] proposed by Vrandečić D built a large number of multi-modal data resources to provide reliable and powerful data sharing query services. In 2017, researchers such as Ferrada S proposed a large-scale link data set IMGpedia [10], which collects a large amount of visual information from the images of the Wikimedia Commons data set, constructs and generates 15 million visual content descriptors, images There are 450 million visual similarities between them, and it is considered a precedent for multimodal data. In 2019, Liu Y and other researchers further advanced the research process of knowledge graphs, and proposed MMKG[11]. MMKG is mainly used to perform relational reasoning by combining different entities and images in different knowledge graphs. MMKG is an entity that contains all entities. A collection of three knowledge graphs of the digital features and (linked to) images, as well as the entity alignment between the knowledge graphs. Therefore, multi-relational link prediction and entity matching communities can benefit from this resource. In the latest multimodal encyclopedia Richpedia [12] proposed by Wang M and other scholars, the multimodality between the image modality London Eye image and the text modality knowledge graph entity (DBpedia entity: London eye) is first constructed. Semantic relationship (rpo:imageof), and then a multi-modal semantic relationship (rpo:nextTo) between the image modality entity London Eye and the image modality entity Big Ben (rpo:nextTo), which greatly promotes multimodal knowledge The construction of the map.

In this article, we take the user's historical click entity as the center, and construct different types of attribute values, entity associations, and pictures, audio, video and other information for the entity (video) to achieve the fusion of different modal information, which is the entity and Different modal information is related to each other, and at the same time, direct interaction and association between different modal information are realized. Based on the constructed multi-modality, I adopted the multi-modal graph neural network structure (MGCN) to extract the information in the multi-modal graph, and based on the fusion features of the extracted multi-modal information, I made personalized recommendations for the target. The advantage of MGCN is that it can use different types of processing methods according to the data types of different nodes in the map to achieve feature extraction. Based on the fusion of these features, it can obtain features with more information. The main contributions of this paper are as follows:

(1) Provides a map construction method based on different modal data, expands the data types of the previous multi-modal map construction, and is the first map to integrate video, picture, audio, and text information.

(2) A new MGNN model was developed. Based on this model, the characteristics of different modal data of different nodes in the map can be extracted to realize the fusion of multi-modal features.

(3) Apply this kind of information to the movie recommendation system, and realize that the performance of the movie recommendation system has been greatly improved compared with the historical performance.

2. Method

In this part, we mainly discuss the current research, including the construction process of the multi-modal knowledge graph, the structure of the MGNN model and the architecture of the recommendation system.

2.1 Multi-modal Knowledge Graphs

Compared with the traditional knowledge graph, the multi-modal knowledge graph introduces more dimensional and morphological information. For the description of an entity, not only uses the text type to describe the associated entities and attributes, but also introduces entity-related images. Information, audio information, etc. Therefore, the problem of information limitation brought by the triple information representation of the traditional map is overcome, and the diversity of information is enriched. Although there are many difficulties in the construction of multimodal maps, there are still many researchers who continue to study and explore them. Some research results have demonstrated the importance of knowledge graphs in tasks such as classification [13,14,15]. From the perspective of knowledge graph representation, knowledge graph representation learning engineering can be divided into entity-based representation methods, feature-based representation methods and event-based representation methods. But our task is mainly applied to video recommendation, and we use an entity-based representation method here.

The multi-modal knowledge graph of this article mainly uses the user's historical click video as the central node. The tail node associated with the central node includes the attribute characteristics of the video, including text characteristics such as actors, directors, and genres, as well as other morphological characteristics. Such as movie tidbits, posters, audio and other information; in addition, it can be extended to other related video nodes through the central node. Its structure is shown in Figure 1.

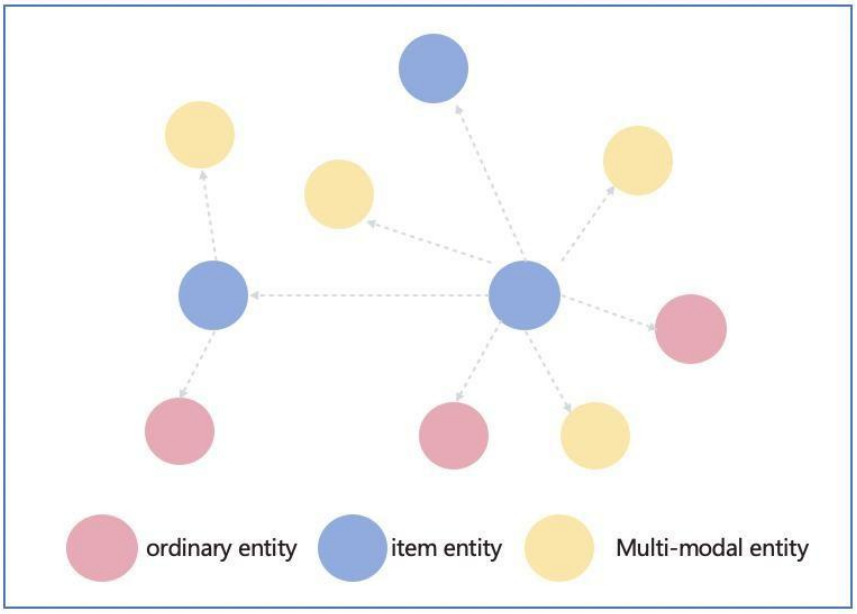


Figure 1. Multi-modal knowledge graph structure diagram

In Figure 1, the blue node represents the item entity node, such as the name of the video that the user has clicked on; the red node represents the traditional node entity, such as the name of the actor and director of the movie in this article; the yellow node represents the multi-modal node, such as Movie

tidbits videos, audios, posters, etc. At the same time, the item video entity clicked by the user can also be associated with other video information clicked by the user connected with it. In this way, the entire multi-modal map spreads out in circles like a "ripple". Whenever it spreads to a new item entity, the map can be used as a new round of diffusion centered on this entity. In this way, a circle of item entities will be diffused with a certain item entity as the center, and then a new circle of items will be diffused with the sub-item as the center, thus forming a multi-hop graph information centered on the central node, each of which Items have their own multi-modal properties. Through such a multi-modal graph, we can not only obtain the multi-modal information of the video that the user recently watched, but also obtain the related multi-hop video information as the recommendation dependent information, which greatly improves the recommendation system's reliance. The completeness of the information.

2.2 MGNN Model

In this part, we mainly introduce the MGNN model we proposed in this article. The MGNN model in this article mainly consists of two parts. The first part is multi-modal feature coding, that is, by judging the data types of different nodes in the multi-modal graph, according to different data types, different coding operations are performed to extract its features vector. The second part is feature fusion and recommendation, that is, by fusing the features acquired in the first part in a certain way, and then recommending based on the fused features.

Multi-modal feature coding is mainly to judge the nature of the corresponding node in the map node, and perform different processing according to different properties.

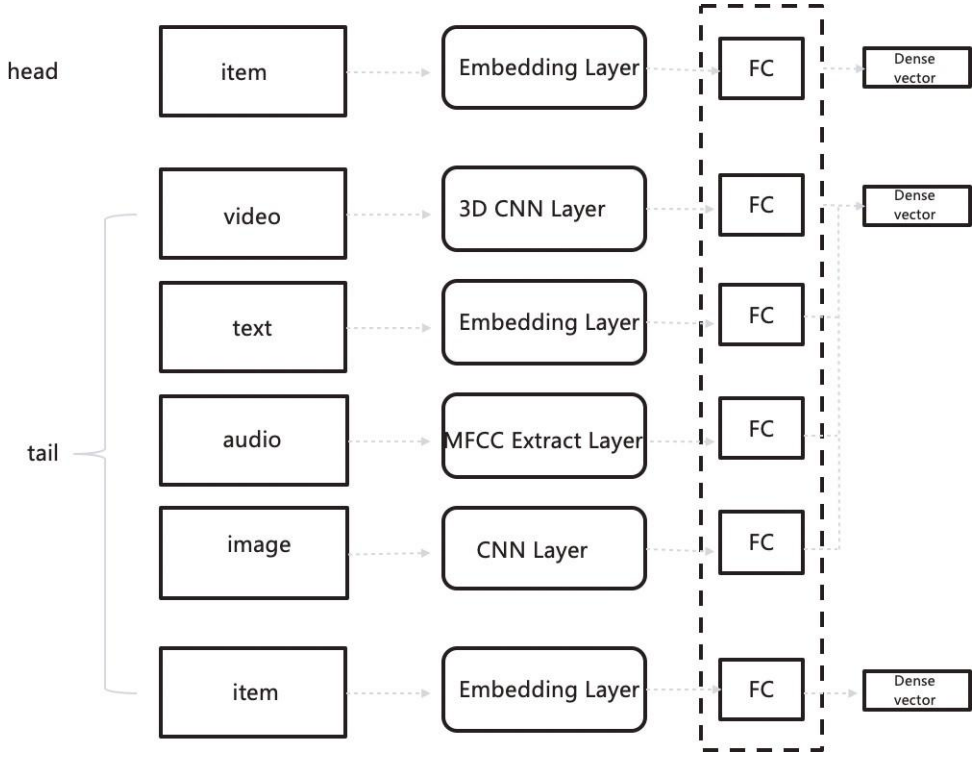


Figure 2. Multi-modal knowledge graph feature coding

The multi-modal feature encoding process, as illustrated in Figure 2, begins by identifying the data type of the node. Depending on the type of data, different feature extraction methods are employed. For text-type nodes, features are encoded using embeddings, as shown in Equation (1):

$$V_h = \sum_n Em(FC(item))$$

Here, EmEmEm represents the embedding operation, itemitemitem denotes the user's historical clicked video (i.e., the feature vector of the item), and FCFCFC refers to the fully connected layer.For video data, features are extracted using a 3D convolutional neural network (3D CNN). Audio data is

transformed into spectrograms and subsequently processed using CNNs for feature extraction. For image data, CNNs are directly utilized for feature extraction, as represented in Equation (2):

$$V_t = \sum_{k=1}^{n} 3DCNN(FC(video)) + Em(FC(text)) + MF(FC(audio)) + CNN(FC(image))$$

In this equation, 3DCNN3DCNN3DCNN represents the 3D convolutional process applied to video data, EmEmEm indicates the embedding operation, MFMFMF refers to Mel spectrum extraction from audio data, CNNCNNCNN denotes feature extraction for image data using convolutional neural networks, and FCFCFC represents the fully connected process.

Next, features from other item nodes linked to the central node are extracted. These item nodes are encoded into feature vectors through an embedding process, as described in Equation (3):

$$V_{ti} = \sum_{n} Em(FC(item))$$

Here, EmEmEm signifies the embedding operation, and FCFCFC represents a fully connected network. The extracted feature vectors are further transformed through a Dense layer to adjust their dimensions, resulting in different feature representations. During this process, the number of hops (KKK) required for feature extraction is determined based on the application's needs. For multi-hop scenarios, the process is repeated for each hop, and the corresponding features are aggregated.

The feature fusion and recommendation module integrates features from various node types, generating composite feature vectors. These fused vectors are then used to retrieve matching items from a material pool. The retrieved materials undergo operations such as rough sorting, fine sorting, and rearrangement. Finally, the top-KKK video materials are selected and recommended to users, completing the recommendation process based on multi-modal knowledge graph features.

As shown in Figure 3, the recommendation module first fuses feature vectors from the multi-modal knowledge graph to generate a unified feature vector. The fusion involves concatenating the vectors and applying a transformation matrix to adjust the dimensionality. The resulting feature vector is then input into the recommendation module, where it retrieves video materials from the library based on the feature codes. The retrieved videos are ranked and refined, ensuring that the final recommended materials are most relevant to the user.

3. Experiments

3.1 Experimental Setup

Since the existing multi-modal data sets are relatively scarce and can not meet the needs of our experiments, we built the data set MMD by ourselves. First of all, we collect a data set, by crawling a large amount of movie website data as the source data, and at the same time extracting entities from the crawled text data, extracting the entities, attributes, relationships, etc., and filling in the map nodes according to the way of map representation , And at the same time search for information such as video highlights, poster pictures, voice clips and other information corresponding to the entity to build our multi-modal knowledge graph. And based on each user’s history, click on the video as a sub-image of the video to mark, where the user interacts with it as a negative example, and the interactive behavior occurs as a positive example, thus generating our labeled data based on the multi-modal map Set, the overall data statistics are shown in Table 1.

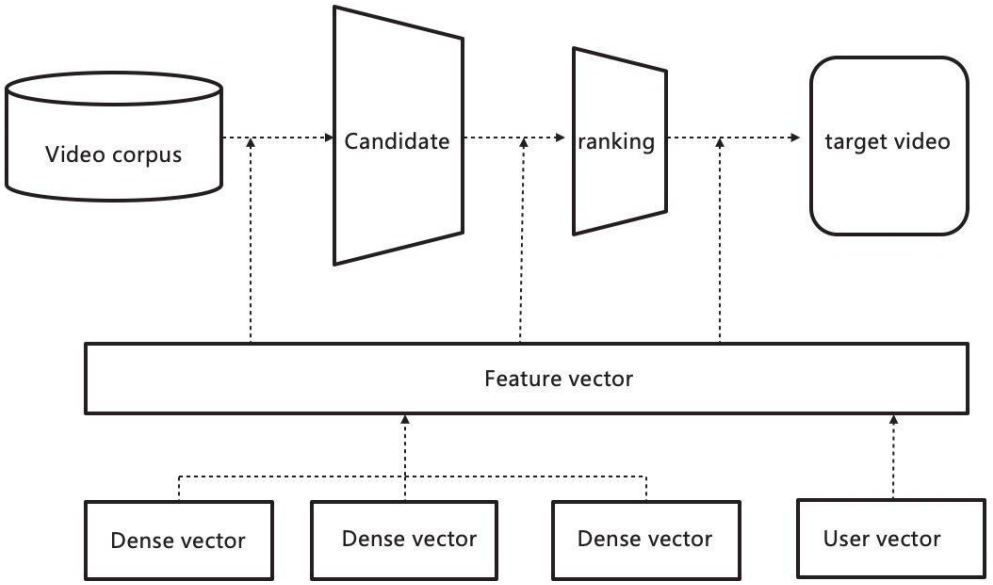


Figure 3. Recommended algorithm structure diagram

Table 1. Statistics of dataset

item	number
# of entities	6580
# of relations	12
# of triplets	14540

On the basis of the above data set, we divided the data set, and randomly selected 20% of the data as the test set to test the performance of the model. During the training process, the remaining data set is divided into training set and validation set according to the ratio of 8:2, which is used for model training and performance verification in the training process. In terms of the evaluation indicators of the model, we have adopted the TOPK method for evaluation. Here, two widely used evaluation matrices are used: recall@k and ndcg@k. The value of k we used here is 10. Finally, we use the average of the evaluation results of all test samples as our result. In terms of model parameters, we adopted the parameter setting method in RippleNet [17]. The optimizer uses the Adam optimizer, the batch size is 256, and the final learning rate is 0.002 through comparison and debugging. For the final hidden vector, this article uses 1024 dimensions.

We have compared the performance between our model and some of the best existing models in the same field. These models mainly include FM-based methods (NFM), knowledge-based mapping methods (RippleNet), and multi-modality-based methods. Knowledge Graph Method (MMGCN).

(1) NFM: NFM is an improvement of the FM algorithm, which mainly proposes an FM model based on neural network [16].

(2) RippleNet: is a recommendation system based on a knowledge graph, which mainly uses the user's clicked product as the central node, and then associates more attributes and products outward, and then uses the network to extract features from the periphery. [17].

(3) MMGCN: MMGCN is the best-performing multi-modal model. It considers the interaction between individual users of each model, where GCN is used to train each part, and finally the information is fused [18].

In addition, in order to verify the impact of the multi-modal graph on the recommendation performance, we also set up our own comparative verification, that is, we filter the data set, generate a graph of plain text nodes, and change the network structure to produce a TGNN network. Compared with the performance of the MGNN network based on the MMD data set in this paper.

3.2 Experimental Result

For comparative experiments between different models, we use our data set as the experimental data set. According to the different models, we have screened and processed the data modalities and forms of the data set, and performed our benchmark model and our model. Training with the same number of steps. Finally, the same test data set is used to test different models. The experimental results are shown in Table 2.

Table 2. Comparison table of model experiment results

Model	recall	ndcg
NFM	0.3321	0.4174
RippleNet	0.3618	0.4232
MMGCN	0.3766	0.4823
MGNN	0.3901	0.5052

The experimental results of all models are shown in the statistical data in Table 2. We can see that the performance effect of the NFM model is the worst, with the recall value reaching 0.3321 and the ndcg value reaching 0.4174; the effect of the RippleNet model is compared with that of the NFM. Great advantage, the recall value reached 0.3618, and the ndcg index reached 0.4232; in addition, the MMGCN model is a model based on a multi-modal knowledge graph. The test results are closest to the results of our MGNN model, with the recall value reaching 0.3901. The ndcg value reached 0.5052; in the end, it was proved through experiments that our model performed best, surpassing the MMGCN second only to our model to a certain extent. The recall value of our model was 0.3901 and the ndcg value was 0.5052.

From table 2, it can be seen that the performance of our MGNN model based on the multi-modal knowledge graph surpasses the best existing models, and the introduction of the multi-modal knowledge graph has greatly improved the performance of the video recommendation algorithm.

In addition, we also conducted a comparative experiment on the model itself. Based on our MMD data set, we filtered and generated a data containing only text nodes. Through the improvement of the multi-modal feature extraction code in the model. We generated a network model TGNN that only extracts the text features in the map, and used it to compare our multi-modal feature network models. The experimental results are shown in Table 3.

Table 3. Comparison table of model experiment results

Model	recall	ndcg
TGNN	0.3507	0.4178
MGNN	0.3901	0.5052

It can be seen from the experimental results in Table 3 that the recall value of the TGNN network based on the simple text node graph is 0.3507, and the ndcg index is 0.4178; while the recall value of the MGNN network based on the multimodal graph is 0.3901 and the ndcg index is 0.5052. Therefore, it can be seen that the performance index of the MGNN model given the multi-modal knowledge graph is higher than that of the TGNN model based on the text graph, so the addition of multi-modal knowledge improves the performance of the recommendation system to a certain extent.

4. Conclusion

In this article, we propose a recommendation model MGNN based on a multi-modal knowledge graph, by introducing the construction process of the multi-modal knowledge graph and the application process of the multi-modal knowledge graph in the recommendation system. In this article, we take

movie recommendation as an example. In order to enhance the auxiliary information and improve the recommendation effect, in the process of constructing the knowledge graph, we have added different data forms such as video, picture, and audio on the basis of the traditional knowledge graph of text nodes. The nodes form a multi-modal knowledge graph. Then this paper uses the proposed MGNN network model to extract features of different forms of data in the multi-modal knowledge graph, and perform feature fusion to form item features for user interaction with extensive information, and then based on these item features and the user's own features Use recommendation algorithm to recommend video materials.

The work of this paper is an exploratory research on the combination of recommendation system and multi-modal knowledge graph. Through our research, it is found that multi-modal knowledge can assist in improving the performance of recommendation to a certain extent. Here, for the representation of multi-modal knowledge graphs, we have studied better graph structure and graph representation. At the same time, based on the data modal types of our graphs, we have further improved the graph neural network and proposed our multi-modal knowledge graph. State graph neural network. In future research, we will further improve the representation and data type of the graph, as well as the structure of the model, hoping to make new breakthroughs in the performance of the recommendation system.

References

[1] Liu, W., Zhang, Z., Li, X., Hu, J., Luo, Y., & Du, J. (2024). Enhancing Recommendation Systems with GNNs and Addressing Over-Smoothing. arXiv preprint arXiv:2412.03097.

[2] Mohsen Jamali and Martin Ester. 2010. A matrix factorization technique with trust propagation for recommendation in social networks. In Proceedings of the 4th ACM conference on Recommender systems. ACM, 135–142.

[3] Yan, X., Jiang, Y., Liu, W., Yi, D., & Wei, J. (2024). Transforming Multidimensional Time Series into Interpretable Event Sequences for Advanced Data Mining. arXiv preprint arXiv:2409.14327.

[4] Luo, Y., Wang, R., Liang, Y., Liang, A., & Liu, W. (2024). Metric Learning for Tag Recommendation: Tackling Data Sparsity and Cold Start Issues. arXiv preprint arXiv:2411.06374.

[5] YuSun,NicholasJingYuan,XingXie,KieranMcDonald,andRuiZhang.2017. Collaborative Intent Prediction with Real-Time Contextual Data. ACM Transac- tions on Information Systems 35, 4 (2017), 30.

[6] Hongwei Wang, Fuzheng Zhang, Jialin Wang, Miao Zhao, Wenjie Li, Xing Xie, and Minyi Guo. 2018. Ripplenet: Propagating user preferences on the knowledge graph for recommender systems. In Proceedings of the 27th ACM International Conference on Information and Knowledge Management. 417–426.

[7] Xiang Wang, Xiangnan He, Yixin Cao, Meng Liu, and Tat-Seng Chua. 2019. Kgat: Knowledge graph attention network for recommendation. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 950–958.

[8] S Auer, C Bizer, G Kobilarov, et al. Dbpedia: A nucleus for a web of open data[M]//The semantic web. Springer, Berlin, Heidelberg, 2007: 722-735.

[9] D Vrandečić, M Krötzsch. Wikidata: a free collaborative knowledgebase[J]. Communications of the ACM, 2014, 57(10): 78-85.

[10] S Ferrada, B Bustos, A Hogan. IMGpedia: a linked dataset with content-based analysis of Wikimedia images[C]//International Semantic Web Conference. Springer, Cham, 2017: 84-93.

[11] Y Liu, H Li, A Garcia-Duran, et al. MMKG: multi-modal knowledge graphs[C]//European Semantic Web Conference. Springer, Cham, 2019: 459-474.

[12] M Wang, G Qi, H F Wang, et al. Richpedia: A Comprehensive Multi-modal Knowledge Graph[C]//Joint International Semantic Technology Conference. Springer, Cham, 2019: 130-145.

[13] Maoxiang Hao, Zhixu Li, Yan Zhao, and Kai Zheng. 2018. Mining High-QualityFine-Grained Type Information from Chinese Online Encyclopedias. In InternationalConference on Web Information Systems Engineering. 345–360.

[14] Hatem Mousselly-Sergieh, Teresa Botschen, Iryna Gurevych, and Stefan Roth.2018. A multimodal translation-based approach for knowledge graph representationlearning. In Proceedings of the Seventh

Joint Conference on Lexical and Computational Semantics. 225–234.

[15] Ruobing Xie, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2016. Imageembodied knowledge representation learning. arXiv preprint arXiv:1609.07028 (2016).

[16] Xiangnan He and Tat-Seng Chua. 2017. Neural factorization machines for sparse predictive analytics. In Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval. 355–364.

[17] Hongwei wang, Fuzheng Zhang, Jialin Wang, et al. "RippleNet: Propagating User Preferences on the Knowledge Graph for Recommender Systems." the 27th ACM International Conference ACM, 2018.

[18] YinweiWei, XiangWang, Liqiang Nie, Xiangnan He, Richang Hong, and Tat-Seng Chua. 2019. MMGCN: Multi-modal Graph Convolution Network for Personalized Recommendation of Micro-video. In Proceedings of the 27th ACM International Conference on Multimedia. 1437–1445.