Multi-Modal KG Convolutional Network for Music Recommender Systems

Subjects: Automation & Control Systems

Contributor: Xiaohui Cui , Xiaolong Qu , Dongmei Li , Yu Yang , Yuxun Li , Xiaoping Zhang

Modern online music services have changed the way people search for and listening to music, offering an extensive array of diverse song catalogues while concurrently enhancing user experiences through personalized optimization. Knowledge graphs (KGs) are a rich source of semantic information for entities and relations, allowing for improved modeling and analysis of entity relations to enhance recommendations.

music information retrieval | multi-modal knowledge graph | graph convolutional networks

1. Introduction

Modern online music services have changed the way people search for and listening to music, offering an extensive array of diverse song catalogues while concurrently enhancing user experiences through personalized optimization [1][2]. In this context, the development of music information retrieval [3] has become crucial in enhancing user experience and improving the profitability of these platforms [4]. Music recommender systems, as the core technology of music information retrieval, can provide personalized music recommendations to users based on their preferences and behavioral patterns, thereby increasing user satisfaction, loyalty, and ultimately promoting revenue growth of the music platform [5][6]. Therefore, the importance of music recommender systems is paramount.

Traditional content-based recommendation methods [4] usually only consider the features of the music itself, neglecting the potential relations between music and other entities, such as artists, albums, and playlists. As a result, they fail to uncover deeper semantic information behind the music [7]. Collaborative filtering (CF) methods [8][9], on the other hand, require a large amount of user behavior data, making them less effective for new users or cold-start problems. Additionally, with the development of mobile internet, the data used for recommendation has become more specific and diverse, including user ratings, music tags, and multi-modal data such as texts, images, audios, and sentiment analysis of the music itself. Therefore, there are still challenges in effectively utilizing side information and multi-modal data to enhance the performance of music recommender systems.

2. Convolutional Neural Networks

In recent years, convolutional neural networks (CNNs) have shown impressive performance in the domains of video [10] and images [11]. However, when it comes to non-Euclidean data structures, such as social networks and knowledge graphs, CNNs' efficacy is limited. To address this issue, researchers have proposed graph convolutional

networks (GCNs), which are an extension of CNNs in the non-Euclidean domain. By integrating the features and labeling information of both the central node and its neighboring nodes, GCNs provide a regular expression form of each node in the graph and input it into CNNs. In this way, GCNs can combine multi-scale information to create higher-level expressions, effectively utilizing both the graph structure information and attribute information. Due to their powerful modeling capabilities, GCNs have found widespread use in recommender systems [12][13]. There are two primary methods for GCNs to perform convolution operations: (1) spectral decomposition graph convolution, which is an eigen decomposition using the Laplacian matrix of the graph, and (2) spatial graph convolution, which leverages the spatial characteristics of graph structure data to explore the representation of neighbor nodes, making the representation of each node's neighboring nodes uniform and regular, which is convenient for convolution operations [14]. KGCN [15] samples the neighbors around a node and dynamically computes local convolutions based on the sampled neighbors to enhance the item embedding representation. LightGCN [16] proposes a lightweight GCN based on an interaction graph by learning users and linearly propagating embedding items on a user–item bipartite graph, and using the weighted sum of embedding item learning across all levels as the final embedding.

MKGCN uses spatial graph convolution for a GCN. While GCNs have been effective in modeling high-order representations of items in KGs, they often neglect modeling user preferences.

3. Multi-Modal Knowledge Graph

Multi-modal knowledge graphs (MMKGs) have become increasingly important in the field of artificial intelligence due to the prevalence of multi-modal data in various domains. MMKGs integrate information from different modalities, such as text, image, and audio, into traditional KGs, which typically only contain structural triples, to improve the performance of downstream KG-related tasks [17]. **Figure 1** illustrates the two main approaches for constructing MMKGs. (Please note that the face images in the figure are sourced from the open-source Generated Faces dataset. The dataset can be accessed via the link: https://generated.photos/datasets#, accessed on 6 June 2023.) The first approach, attribute-based MMKGs, considers multi-modal data as specific attribute values of entities or concepts, such as the "poster" and "audio" of a music entity in **Figure 1**a. The second approach, entity-based MMKGs, treats multi-modal data as separate entities in the KG, as shown by the image-based representation of singer and country entities in **Figure 1**b. However, entity-based MMKGs do not fuse multi-modal data and therefore limit the exploitation of multi-modal information [18][19].

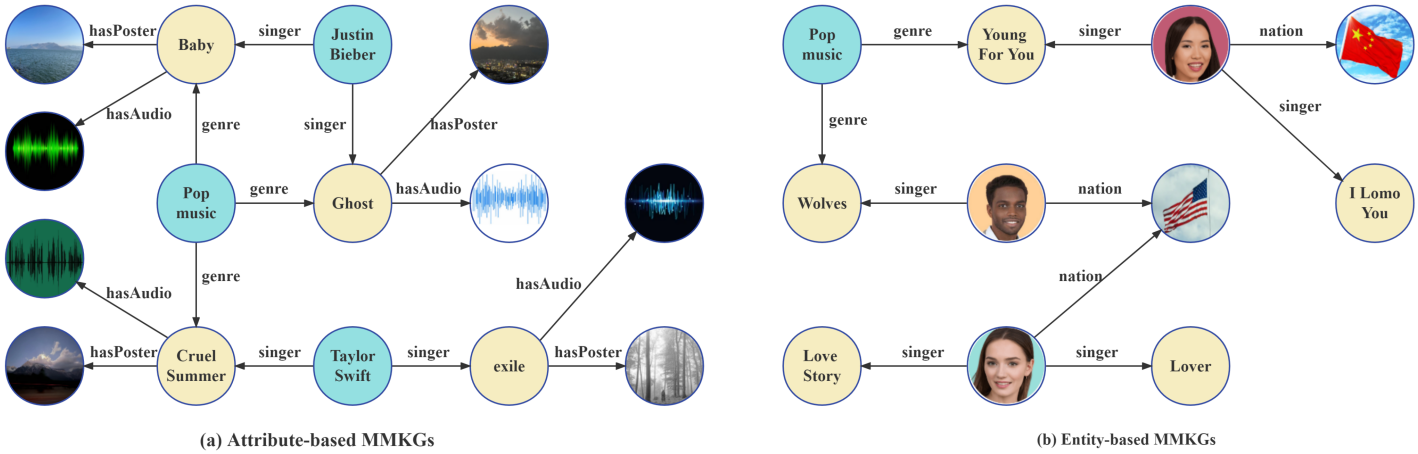

Figure 1. Illustration of the multi-modal knowledge graph. The left subfigure shows an illustration of an attribute-based MMKG and the right subfigure shows an illustration of an entity-based MMKG.

4. Recommendations with MMKGs

As MMKGs are a relatively new concept, there is limited related work on MMKG-based recommender systems. Researchers propose three classifications of existing MMKG-based recommender systems from the perspective of multi-modal feature fusion [20]: (1) The feature-fusion method, also known as the early-fusion method, concatenates features extracted from different modalities into a single high-dimensional feature vector that is then fed into a downstream task. For instance, MKGAT [21] first performs separate feature representations of multi-modal data such as text, image, and triples, and then aggregates the embedding representation of the feature vectors from each modality to make recommendations. However, this method is limited in its ability to model complex relations between modalities. (2) The result-fusion method, also known as the post-fusion method, obtains decisions based on each modality and then integrates these decisions by applying algebraic combination rules for multiple prediction class labels (e.g., maximum, minimum, sum, mean, etc.) to obtain the final result. For example, MMGCN [22] is based on a knowledge graph of three different modalities (text, image, and audio) and then performs user–item interaction predictions on all three knowledge graphs simultaneously. It then linearly aggregates the prediction scores for each modality to obtain the final prediction score. However, this method cannot capture the interconnections between different modalities and requires corresponding multi-modal data for each item. (3) The model-fusion method is a deeper fusion method that produces more optimal joint discriminative feature representations for classification and regression tasks. For instance, MKRLN [23] generates path representations by combining structural and visual information of entities and incorporates the idea of reinforcement learning to iteratively introduce visual features to determine the next step in path selection.

References

1. Hagen, A.N. The playlist experience: Personal playlists in music streaming services. Pop. Music. Soc. 2015, 38, 625–645.

2. Kamehkhosh, I.; Bonnin, G.; Jannach, D. Effects of recommendations on the playlist creation behavior of users. User Model. User Adapt. Interact. 2020, 30, 285–322.
3. Burgoyne, J.A.; Fujinaga, I.; Downie, J.S. Music information retrieval. In A New Companion to Digital Humanities; Wiley: Hoboken, NJ, USA, 2015; pp. 213–228.
4. Murthy, Y.V.S.; Koolagudi, S.G. Content-based music information retrieval (cb-mir) and its applications toward the music industry: A review. ACM Comput. Surv. CSUR 2018, 51, 1–46.
5. Schedl, M.; Zamani, H.; Chen, C.W.; Deldjoo, Y.; Elahi, M. Current challenges and visions in music recommender systems research. Int. J. Multimed. Inf. Retr. 2018, 7, 95–116.
6. Schedl, M.; Gómez, E.; Urbano, J. Music information retrieval: Recent developments and applications. Found. Trends Inf. Retr. 2014, 8, 127–261.
7. Wu, L.; He, X.; Wang, X.; Zhang, K.; Wang, M. A survey on accuracy-oriented neural recommendation: From collaborative filtering to information-rich recommendation. IEEE Trans. Knowl. Data Eng. 2022, 35, 4425–4445.
8. Wang, X.; He, X.; Wang, M.; Feng, F.; Chua, T.S. Neural graph collaborative filtering. In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, Paris, France, 21–25 July 2019; pp. 165–174.
9. Zhang, H.R.; Min, F.; Zhang, Z.H.; Wang, S. Efficient collaborative filtering recommendations with multi-channel feature vectors. Int. J. Mach. Learn. Cybern. 2019, 10, 1165–1172.
10. Guo, J.; Han, K.; Wu, H.; Tang, Y.; Chen, X.; Wang, Y.; Xu, C. Cmt: Convolutional neural networks meet vision transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 21–24 June 2022; pp. 12175–12185.
11. Sun, Y.; Xue, B.; Zhang, M.; Yen, G.G.; Lv, J. Automatically designing CNN architectures using the genetic algorithm for image classification. IEEE Trans. Cybern. 2020, 50, 3840–3854.
12. Hamilton, W.; Ying, Z.; Leskovec, J. Inductive representation learning on large graphs. In Proceedings of the 2017 Annual Conference on Neural Information Processing Systems: Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 30.
13. Ying, R.; He, R.; Chen, K.; Eksombatchai, P.; Hamilton, W.L.; Leskovec, J. Graph convolutional neural networks for web-scale recommender systems. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, London, UK, 19–23 August 2018; pp. 974–983.
14. Bruna, J.; Zaremba, W.; Szlam, A.; LeCun, Y. Spectral networks and deep locally connected networks on graphs. In Proceedings of the 2nd International Conference on Learning Representations (ICLR 2014), Banff, AB, Canada, 14–16 April 2014.

15. Wang, H.; Zhao, M.; Xie, X.; Li, W.; Guo, M. Knowledge graph convolutional networks for recommender systems. In Proceedings of the World Wide Web Conference, San Francisco, CA, USA, 13–17 May 2019; pp. 3307–3313.
16. He, X.; Deng, K.; Wang, X.; Li, Y.; Zhang, Y.; Wang, M. Lightgcn: Simplifying and powering graph convolution network for recommendation. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, Online, 25–30 July 2020; pp. 639–648.
17. Zhu, X.; Li, Z.; Wang, X.; Jiang, X.; Sun, P.; Wang, X.; Xiao, Y.; Yuan, N.J. Multi-modal knowledge graph construction and application: A survey. IEEE Trans. Knowl. Data Eng. 2022, 1, 1–20.
18. Mousselly-Sergieh, H.; Botschen, T.; Gurevych, I.; Roth, S. A multimodal translation-based approach for knowledge graph representation learning. In Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics, New Orleans, LA, USA, 5–6 June 2018; pp. 225–234.
19. Pezeshkpour, P.; Chen, L.; Singh, S. Embedding Multimodal Relational Data for Knowledge Base Completion. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp. 3208–3218.
20. Guo, W.; Wang, J.; Wang, S. Deep multimodal representation learning: A survey. IEEE Access 2019, 7, 63373–63394.
21. Sun, R.; Cao, X.; Zhao, Y.; Wan, J.; Zhou, K.; Zhang, F.; Wang, Z.; Zheng, K. Multi-modal knowledge graphs for recommender systems. In Proceedings of the 29th ACM International Conference on Information & Knowledge Management, Online, 19–23 October 2020; pp. 1405–1414.
22. Wei, Y.; Wang, X.; Nie, L.; He, X.; Hong, R.; Chua, T.S. MMGCN: Multi-modal graph convolution network for personalized recommendation of micro-video. In Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 21–25 October 2019; pp. 1437–1445.
23. Tao, S.; Qiu, R.; Ping, Y.; Ma, H. Multi-modal knowledge-aware reinforcement learning network for explainable recommendation. Knowl.-Based Syst. 2021, 227, 107217.