Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 17, No. 3, 2026

Recommendation Algorithm Fusing Multimodal Auxiliary Information for Multi-view Social Recommendation

Tiantian Liu *, Haiyan Zi

Sias University, Zhengzhou, Henan, 451150, China

*** Corresponding author:** Tiantian Liu

Abstract: Aiming at the problems of data sparsity, cold start and the limitation of single-modal auxiliary information in recommendation systems, this paper proposes a multi-view social recommendation algorithm that fuses multimodal auxiliary information. The algorithm consists of a user-consistent social recommendation module and a multimodal information fusion module. The former dynamically measures the preference similarity between the target user and social neighbors with an attention mechanism, selects preference-consistent neighbors through a differentiable gate for neighborhood aggregation, and extracts item features from both the knowledge graph view and the user-item interaction view, which are integrated by a cross-attention fusion unit. The latter encodes user raw features into low-dimensional dense vectors with multi-layer perceptrons, performs multi-layer information propagation with residual connections, and adopts attention-based late fusion for textual and visual modalities so that modality weights adapt to users and scenarios dynamically. Experiments are conducted on two public datasets, Yelp Open and LastFM-2k, compared with six baselines (BPR-MF, NeuMF, LightGCN, SocialMF, MMGCN, VBPR) using Recall, NDCG and Precision as evaluation metrics. Experimental results show that the proposed algorithm achieves the best performance on both datasets, and ablation studies verify the effectiveness of user consistency filtering, social aggregation, multi-view item feature extraction and dynamic multimodal fusion.

Keywords: Social Recommendation; Multimodal Fusion; Attention Mechanism.

1. Introduction

The rapid growth of the Internet has greatly enriched platform contents while making user interests vague, which leads to the problem of information overload. Recommender systems have become one of the most effective means to alleviate information overload, whose core value is to build accurate matching between users and items[1].

Collaborative filtering(CF)[2, 3] is the most widely used recommendation technique, which exploits historical user-item interactions and common preferences for recommendation. However, CF suffers from data sparsity, cold start and poor scalability: when user-item interactions are insufficient or unevenly distributed, the algorithm cannot fully capture real user preferences; for new users and new items, the lack of historical interactions makes it difficult to generate effective personalized recommendations. To alleviate these problems, researchers have introduced auxiliary information such as social relations, knowledge graphs and multimodal information.

In recent years, incorporating social relations into recommender systems has attracted wide attention. Wu et al. proposed SocialGCN [4], which designs a deep model to capture the distinctive features of social networks for user and item modeling. Yu et al. proposed a self-supervised multi-channel hypergraph convolutional network[5], which models high-order user-item and user-user relations with hypergraph convolution and self-supervised learning to obtain more comprehensive user representations. However, most existing social recommendation models neglect the inconsistency of users: not every social neighbor contributes equally to the recommendation of the target user, and aggregating information from inconsistent neighbors may weaken the performance of recommender systems.

Meanwhile, relying only on a single modality of auxiliary information limits the comprehensive understanding of user needs and item characteristics. Inspired by the multi-sensory perception mechanism of human beings, different modalities describe the same object from different perspectives, and integrating complementary multimodal information can restore the essence of the object more comprehensively[6]. For example, movie recommendation can emphasize poster image features while product recommendation relies more on descriptive texts; fixed-weight modality fusion cannot adapt to such scenario differences.

Therefore, aiming at the problems that traditional recommendation algorithms ignore the inconsistency of social neighbors and use only single-modal auxiliary information, this paper proposes a multi-view social recommendation algorithm fusing multimodal auxiliary information (Recommendation Algorithm Fusing Multimodal Auxiliary Information for Multi-view Social Recommendation, RAMSR). The main contributions are summarized as follows:

(1) We propose a user-consistent social modeling method. An attention mechanism dynamically measures the preference similarity between the target user and social neighbors, and a differentiable gate selects consistent neighbors for neighborhood aggregation, together with an attention consistency regularization, alleviating the inaccurate feature representation caused by ignoring user preference differences.

(2) We propose a multi-view item feature extraction method. Item features are independently learned from the knowledge graph view (item attributes and entity relations) and the user-item interaction view, and integrated by a cross-

attention fusion unit, breaking through the limitation of single-view modeling.

(3) We propose an attention-driven multimodal dynamic fusion method. Attention-based late fusion is adopted for textual and visual modalities on top of multi-layer information propagation with residual connections, so that modality weights adapt to different recommendation scenarios automatically.

We conduct comparative experiments with six baselines on two public datasets, and design ablation experiments to verify the effectiveness of each component.

2. Related Work

2.1. Social Recommendation

With the popularity of social platforms, recommender systems incorporating social relations have attracted much attention. Early studies mainly adopted matrix factorization based methods, such as SoRec [7] which jointly factors the user-item rating matrix and the user social relations, and TrustMF [8] which performs matrix factorization based on trust propagation. Recently, many studies transform user relations and historical interactions into graph structures and adopt graph neural networks to learn node representations, such as SocialGCN, GraphRec[9] and DiffNet++[10]. However, most of these methods treat all social neighbors equally and ignore the preference inconsistency among users.

2.2. Knowledge Graph Based Recommendation

Knowledge-graph-based recommendation can be divided into path-based, embedding-based and hybrid methods. Knowledge graph embedding (KGE)[11] based methods have become increasingly popular due to their simplicity and effectiveness, e.g., CKE jointly learns CF and KGE, and KGAT [12] recursively propagates entity embeddings on the knowledge graph. However, most of these methods utilize item information from a single view and fail to fully exploit the multi-view information of item attributes and historical interactions.

2.3. Multimodal Recommendation

Multimodal recommendation aims at fusing textual, visual and other modalities to improve recommendation performance. Related studies usually extract visual features with pre-trained networks (e.g., VGG, ResNet) and extract textual features with bag-of-words or pre-trained language models, and then fuse them with concatenation, gating or attention mechanisms. Representative works include VBPR [13], which introduces pre-trained visual features into Bayesian personalized ranking to capture users' visual preferences, and MMGCN [14], which constructs separate propagation graphs for textual and visual modalities and fuses multimodal features. Fusion strategies can be categorized into early fusion, late fusion and hybrid fusion. Late fusion preserves the characteristics of each modality better since each modality is modeled independently before fusion. This paper adopts attention-based late fusion so that modality weights adapt to users and scenarios dynamically.

3. Methodology

3.1. Notation and Problem Definition

In a typical social recommendation scenario, let $U=\{u_1,u_2,\cdots,u_m\}$ and $I=\{i_1,i_2,\cdots,i_n\}$ denote the user set and the item set respectively, where m and n are the numbers of users and items. The user-item interaction matrix $R\in\{0,1\}^{\{m,n\}}$ is constructed from historical interactions, where $r_{ui}=1$ indicates that user u has interacted with item i (e.g., click, play or listen), and $r_{ui}=0$ otherwise. The knowledge graph is represented by triples (h,r,t), where h and t are entities and r is the relation. The social graph is denoted as $G=(U,E)$, where an edge $(u,v)\in E$ indicates a social relation between users u and v. Each item i is associated with a textual feature vector x_i^t and a visual feature vector x_i^v, and each user u is associated with a raw feature vector x_u (e.g., demographic or profile features). Given the above inputs, the goal is to learn a prediction function $\hat{y}_{ui}=f(u,i,\Theta)$ that estimates the probability that user u will interact with an unseen item i, where Θ denotes the model parameters.

3.2. Framework Overview

The proposed algorithm consists of two parts: the user-consistent social recommendation module (SAUM) and the multimodal information fusion module (RAMA). The framework is illustrated in Fig.1. In the SAUM module, user modeling selects consistent social neighbors with the attention mechanism, aggregates their features to obtain the social-space user representation, and combines user-item interaction information to obtain the complete user representation; item modeling extracts item features from both the knowledge graph view and the interaction view, which are integrated by a cross-attention fusion unit. In the RAMA module, user raw features are deeply analyzed with multi-layer perceptrons, textual and visual modalities are modeled separately, and attention-based late fusion is performed on top of multi-layer information propagation. Finally, the inner product of the user representation and the item representation produces the predicted score.

3.3. User-Consistent Social Modeling

This module uses the attention mechanism to make the target user consistent with neighbors in the social graph, selects consistent neighbors through a differentiable gate, aggregates their features to obtain the social-space user representation, and introduces a consistency regularization to align attention weights with neighbor preference similarity.

3.3.1. Consistent Neighbor Selection

To overcome the user inconsistency problem, the relevance score between the target user and each neighbor is computed first. The attention scoring function is a multi-layer perceptron (MLP) taking the linearly transformed user embedding e_u and neighbor embedding e_v as input:

$$s(u,v)=w_a^T\cdot ReLU(W_u\cdot e_u+W_v\cdot e_v+b_a)$$

where W_u, W_v and b_a are learnable parameters and ReLU is the activation function. The attention weights are obtained by normalizing the scores with the softmax function:

$$\alpha(u,v)=exp(s(u,v))/\sum_{v'\in N(u)}exp(s(u,v'))$$

where $N(u)$ is the neighbor set of user u. Then a

differentiable sigmoid gate is adopted to filter consistent neighbors, keeping neighbors whose attention weights are above the threshold and suppressing inconsistent ones:

$$g(u,v)=\sigma\left(\left(\alpha(u,v)-\tau/d_u\right)/T\right)$$

where τ is the consistency threshold coefficient, d_u is the number of valid neighbors of user u, T is the gate temperature and σ is the sigmoid function. The neighbor with the largest attention weight is always kept to guarantee a non-empty set. The filtered attention weights are re-normalized to obtain $\alpha_{f(u,v)}$. The threshold τ controls the filtering strength: a larger τ keeps fewer neighbors with purer information but may lose effective information, while a smaller τ keeps more neighbors but may introduce interfering neighbors. The selection of τ balances filtering effectiveness and information sufficiency.

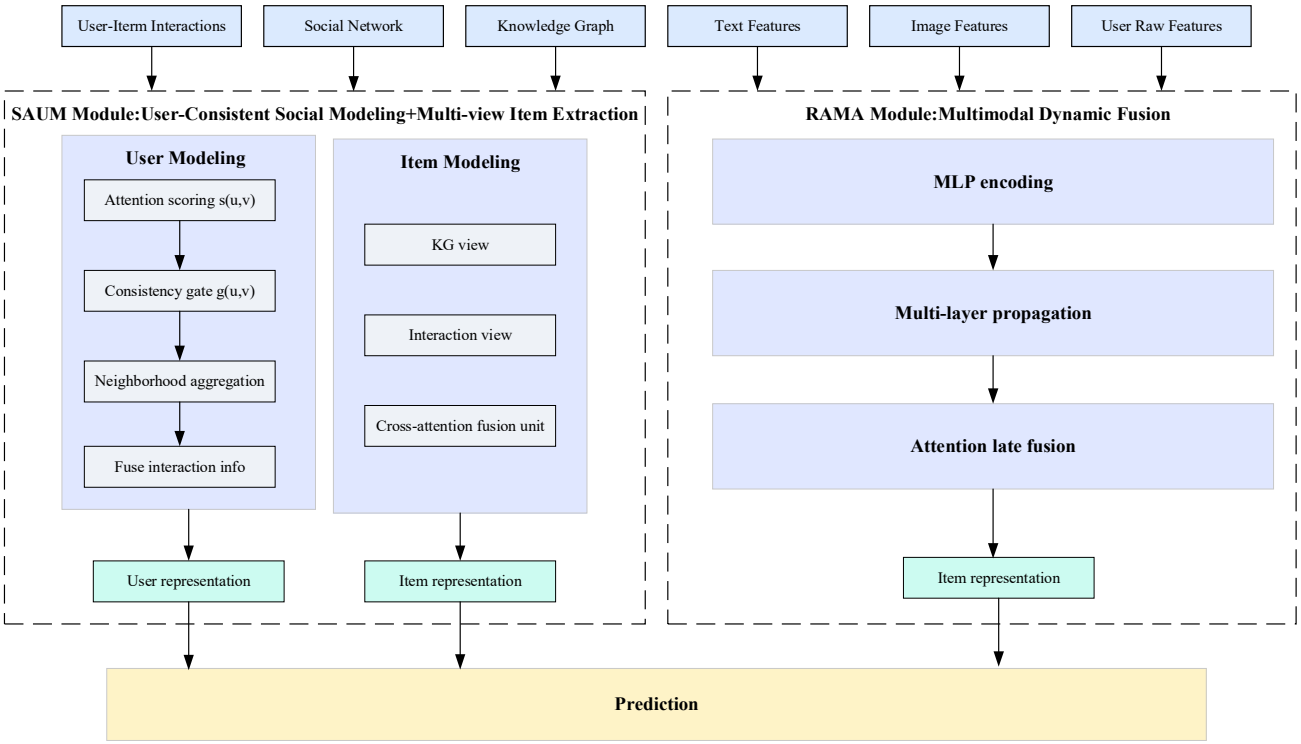


Fig 1. Framework of the proposed recommendation algorithm

3.3.2. Social Neighborhood Aggregation

The features of filtered consistent neighbors are aggregated in a weighted manner to obtain the social-space user representation:

$$e_u^{\ S}=\sum_{v\in N(u)}\alpha_{f(u,v)}\cdot e_v$$

The target user embedding and its social representation are combined to obtain the social-enhanced user representation:

$$e_u=e_u^{\ 0}+e_u^{\ S}$$

3.3.3. Attention Consistency Regularization

To make the attention weights accurately reflect the preference similarity between neighbors and the target user, a consistency regularization is introduced to encourage the alignment between attention weights and the cosine similarity of user embeddings:

$$L_{aux}=\left(\frac{1}{|U'|}\right)\sum_{u\in U'}\frac{1}{|N(u)|}\sum_{u\in N(u)}\alpha(u,v)\cdot\left(1-cos\left(e_u,e_v\right)\right)$$

where U' is the set of sampled users in a training batch and cos is the cosine similarity function. This regularization enables the attention mechanism to learn the mapping from preference similarity to attention weights, making the consistency filtering threshold well-targeted.

3.4. Multi-view Item Feature Extraction

This module learns item representations from the knowledge graph view and the user-item interaction view, and integrates them through a cross-attention fusion unit. The knowledge graph contains item attributes and entity relations, while the interaction view reflects the diverse opinions of different users on the same item.

3.4.1. Knowledge Graph View

The attribute entities associated with item i (e.g., movie genres and keywords, artist tags) are treated as neighbors of the knowledge graph view. A self-anchored attention mechanism is adopted: the item embedding serves as the anchor and is concatenated with entity embeddings as the input of the attention network, producing entity attention weights $\beta(i,e)$. The knowledge graph view feature is obtained by weighted aggregation:

$$e_i^{\ K}=\sum_{e\in E(i)}\beta(i,e)\cdot e_e$$

where $E(i)$ is the entity set of item i and e_e is the entity embedding. The self-anchored mechanism guarantees that the view feature of any item (including cold-start items) is non-empty.

3.4.2. Interaction View

The historical interactive users of item i are treated as neighbors of the interaction view. With the same self-anchored attention mechanism, the interaction view feature is obtained:

$$e_i^{\ U}=\sum_{u\in U(i)}\gamma(i,u)\cdot e_u$$

where $U(i)$ is the set of users who interacted with item i and $\gamma(i,u)$ is the interactive-user attention weight.

3.4.3. Cross-attention Fusion Unit

To fully integrate the knowledge graph view feature and the interaction view feature, a cross-attention fusion unit is designed. Taking the knowledge graph feature as the query and the interaction feature as the key/value:

$$g_1 = \sigma\left(\frac{\langle W_q \cdot e_i^K, W_k \cdot e_i^U \rangle}{\sqrt{d}}\right)$$

$$f^K = e_i^K + g_1 \cdot \left(W_v \cdot e_i^U\right)$$

where $<\ >$ denotes the inner product and d is the embedding dimension. Symmetrically, f^U is obtained by querying the knowledge graph feature with the interaction feature. The final item representation is produced by fusing the two outputs with a nonlinear transformation:

$$e_i = ReLU\left(W_f \cdot \left[f^K; f^U\right] + b_f\right)$$

3.5. Multimodal Dynamic Fusion

This module focuses on textual and visual modalities. User raw features are deeply analyzed with multi-layer perceptrons to obtain low-dimensional dense vectors; multi-layer information propagation models high-order correlations; attention-based late fusion dynamically integrates multimodal information.

3.5.1. Modality Feature Encoding

User raw features, item textual features and item visual features are encoded with separate multi-layer perceptrons:

$$h_u^0 = MLP_u\left(x_u\right)$$

$$h_t^0 = MLP_t\left(x_t\right)$$

$$h_v^0 = MLP_v\left(x_v\right)$$

where x_u, x_t and x_v are the user raw feature, item textual feature and item visual feature respectively, and MLP_u, MLP_t, MLP_v are two-layer fully connected networks.

3.5.2. Multi-layer Information Propagation

Based on the user-item interaction graph, a symmetrically normalized adjacency matrix A is constructed. On the user side, information from both textual and visual channels is aggregated; on the item side, each channel propagates independently. Residual connections are introduced to alleviate the gradient vanishing problem of deep networks:

$$h_u^l = h_u^{l-1} + A \cdot \left(h_t^{l-1} + h_v^{l-1}\right)$$

$$h_t^l = h_t^{l-1} + A^T \cdot h_u^{l-1}$$

$$h_v^l = h_v^{l-1} + A^T \cdot h_u^{l-1}$$

After L layers of propagation, the representations of all layers are summed to obtain the final user representation and per-modality item representations:

$$h_u = \sum_l h_u^l$$

$$h_t = \sum_l h_t^l$$

$$h_v = \sum_l h_v^l$$

3.5.3. Attention-based Late Fusion

To adapt to different recommendation scenarios (e.g., movie recommendation emphasizing images while product recommendation emphasizing texts), attention-based late fusion is performed for textual and visual modalities conditioned on the target user representation h_u. The δ-parameterization is adopted so that the attention weights start from the uniform fusion (0.5/0.5) and adapt during training:

$$\delta = tanh\left(w_g^T \cdot ReLU\left(W_g \cdot \left[h_u; h_t; h_v\right]\right)\right) \cdot 0.5$$

$$a_t = 0.5 + \delta$$

$$a_v = 0.5 - \delta$$

$$h_i = a_t \cdot h_t + a_v \cdot h_v$$

where $[\cdot;\cdot]$ denotes the concatenation operation. The δ-parameterization guarantees that the complete model starts from the uniform-fusion solution, i.e., its performance is not lower than that of fixed-weight fusion.

3.6. Prediction and Loss Function

For the SAUM module, the predicted score is the inner product of the user representation and the item representation: $\hat{y}_{ui} = \langle e_u, e_i \rangle$; for the RAMA module, $\hat{y}_{ui} = \langle h_u, h_i \rangle$. The model is trained with the implicit-feedback binary cross-entropy loss and negative sampling:

$$L = -\sum_{(u,i)\in\Omega^+} log\,\sigma\left(\hat{y}_{ui}\right) - \sum_{(u,i)\in\Omega^-} log\,\sigma\left(-\hat{y}_{ui}\right) + \lambda_1 \cdot L_{aux} + \lambda_2 \cdot \left\|\Theta\right\|^2$$

where Ω^+ and Ω^- are the positive and negative sample sets, σ is the sigmoid function, λ_1 is the consistency regularization weight, λ_2 is the regularization coefficient, and Θ denotes the model parameters.

4. Experiments

4.1. Datasets

Two public datasets are adopted in the experiments, and their statistics are summarized in Table 1.

(1) Yelp Open: a business review dataset provided by Yelp Inc. for public competition. We select active users and their friends from three cities, Philadelphia, Tucson and Tampa, forming a subset containing 407,691 review records of 18,524 users on 20,000 businesses, together with 11,583 real friend relations (social network). The textual modality uses business category TF-IDF features (256 dimensions); the visual modality uses business photos fed into ResNet-18 to extract 512-dimensional features; the entities of the knowledge graph view are business categories. The social relations of this dataset are real but relatively sparse, which can be used to verify the robustness of the proposed algorithm in sparse social scenarios.

(2) LastFM-2k: a music listening dataset from HetRec 2011, containing 92,834 listening records of 2100 users on 18,745 artists, 25,434 friend relations (real social network) and artist tag information. The textual modality uses artist tag TF-IDF features (256 dimensions); the visual modality uses album covers fed into ResNet-18 to extract 512-dimensional features; the entities of the knowledge graph view are artist tags.

4.2. Evaluation Metrics

Three widely used metrics for social recommendation are

adopted: Recall, NDCG and Precision. In the test phase, all test positive samples of each user are mixed with 99 sampled negative samples for ranking. The metrics are defined as follows:

$$Recall@K = \frac{\left|\{i \in top-K : r_{ui} = 1\}\right|}{\left|\{i : r_{ui} = 1\}\right|}$$

Table 1. The statistics of the datasets

Dataset	Users	Items	Interactions	Training	Social links	Entities
Yelp Open	18524	20000	407691	377859	11583	566
LastFM-2K	2100	18745	92834	89068	25434	3686

$$NDCG@K = \frac{DCG@K}{IDCG@K}$$

$$Precision@K = \frac{\left|\{i \in top-K : r_{ui} = 1\}\right|}{K}$$

where $DCG@K = \sum_{i \in top-K} \frac{2^{r_{ui}} - 1}{log_2(pos_i + 1)}$, pos_i is the position of the relevant item in the ranking, and IDCG@K is the DCG value in the ideal ranking.

4.3. Baselines

To verify the effectiveness of the proposed algorithm, it is compared with six representative baselines:

(1) BPR-MF [15]: Bayesian personalized ranking matrix factorization based on implicit feedback, using only user-item interactions.

(2) NeuMF [16]: neural collaborative filtering, fusing generalized matrix factorization and multi-layer perceptrons, using only interactions.

(3) LightGCN[17]: a simplified GCN that keeps only the neighborhood aggregation part and propagates on the user-item bipartite graph.

(4) SocialMF[18]: a classic social recommendation baseline that introduces social-neighbor consistency regularization on the basis of matrix factorization.

(5) VBPR [13]: visual Bayesian personalized ranking that incorporates pre-trained visual features into matrix factorization to capture users' visual preferences on items.

(6) MMGCN [14]: a multimodal graph convolution network that constructs separate propagation graphs for textual and visual modalities and fuses multimodal features for recommendation.

4.4. Experimental Settings

The experiments are implemented with Python 3.12 and PyTorch 2.13. The data are split by time: the last interaction of each user is used as the test set, the penultimate interaction as the validation set, and the remaining interactions as the training set. The embedding dimension is 64, the learning rate is 0.001, the batch size is 2048, the number of negative samples per positive sample is 4, the regularization coefficient is 0.00001, the consistency neighbor filtering threshold is τ=0.5, the consistency regularization weight is λ_1=0.5, the number of propagation layers is L=2, and the number of training epochs is 60 with early stopping based on the validation NDCG@10.

4.5. Comparison Results and Analysis

Comparative experiments are conducted on Yelp Open and LastFM-2k, and the results are shown in Table 2 and Fig.2.

Table 2. Comparative experimental results

Model	Yelp Open				LastFM-2K			
	N@10	P@10	N@20	P@20	N@10	P@10	N@20	P@20
BPR-MF	0.3385	0.0568	0.3768	0.0360	0.5779	0.0757	0.5909	0.0405
NeuMF	0.4558	0.0712	0.4882	0.0420	0.5815	0.0769	0.5950	0.0411
LightGCN	0.3794	0.0623	0.4219	0.0395	0.5211	0.0728	0.5414	0.0404
SocialMF	0.3122	0.0524	0.3512	0.0339	0.5031	0.0704	0.5209	0.0387
VBPR	0.3154	0.0523	0.3520	0.0334	0.5846	0.0756	0.6009	0.0411
MMGCN	0.4845	0.0748	0.5142	0.0432	0.6589	0.0820	0.6698	0.0432
RAMSR	0.5663	0.0805	0.5910	0.0451	0.6722	0.0825	0.6825	0.0433

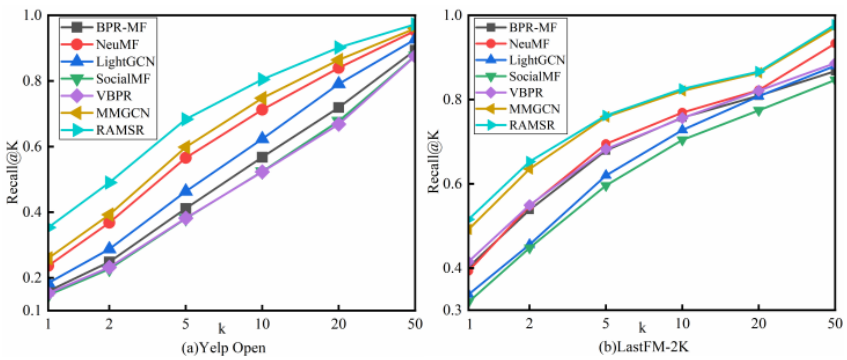


Fig 2. Comparison of Recall@k results of different models

(1)On Yelp Open, where the social relations are real but extremely sparse (users have 0.63 friends on average) and the interactions are limited, the complete model RAMSR proposed in this paper still achieves the best performance, reaching Recall@10, NDCG@10 and Precision@10 of 0.8046, 0.5663 and 0.0805 respectively, with relative improvements of 7.6%, 16.9% and 7.6% over the best baseline MMGCN, and a relative improvement of 53.6% in Recall@10 over the classical social recommendation baseline SocialMF. This indicates that the proposed algorithm can effectively exploit user-consistent social information and multimodal complementary information even when social relations are extremely sparse, demonstrating strong robustness to sparse social networks.

(2)On LastFM-2k, which has a relatively dense social network (12.1 friends per user on average) and contains multimodal information such as artist tags and album covers, RAMSR also achieves the best performance, reaching Recall@10, NDCG@10 and Precision@10 of 0.8253, 0.6722 and 0.0825 respectively, with relative improvements of 0.6%, 2.0% and 0.6% over the strongest baseline MMGCN, and a relative improvement of 17.2% in Recall@10 over the social baseline SocialMF. This also shows that on datasets with dense social relations and relatively balanced modality signals, the multimodal graph convolution baseline MMGCN is already competitive, while the proposed algorithm still obtains further gains through user-consistent social modeling and dynamic multimodal fusion.

(3)Taking the two datasets together, RAMSR achieves the best performance on both: the advantage is significant on Yelp Open with sparse interactions and extremely sparse social relations (7.6% improvement in Recall@10 over the best baseline MMGCN), while it remains leading on LastFM-2k with dense social relations (0.6% improvement over the best baseline MMGCN); compared with the traditional social recommendation baseline SocialMF, RAMSR improves Recall@10 by 53.6% and 17.2% on the two datasets respectively. This indicates that the overall design of "user-consistent social modeling + multi-view item feature extraction + dynamic multimodal fusion" is advantageous on datasets with different social densities and modality distributions, verifying the effectiveness and robustness of the proposed algorithm.

4.6. Ablation Study

To evaluate the contribution of each component, taking the complete model RAMSR as the core, ablation experiments are designed: on the basis of the complete model, we remove the consistency neighbor filtering (τ =0), remove the social aggregation, keep only the interaction view (remove the knowledge graph view), fix the modality fusion weights to 0.5/0.5, and remove the multi-layer information propagation (L =0) respectively. The results are shown in Table 3.

Table 3. Ablation results of RAMSR (Recall@10)

Variant	Yelp Open	LastFM-2K
RAMSR	0.8046	0.8253
w/o consistency filtering	0.8039	0.8216
w/o social aggregation	0.8014	0.8253
interaction view only	0.7733	0.8237
fixed-weight fusion	0.7985	0.8290
without propagation	0.8015	0.8072

As shown in Table 3: (1) Removing the consistency neighbor filtering decreases Recall@10 by 0.1% and 0.5% on Yelp Open and LastFM-2k respectively, indicating that filtering out preference-inconsistent neighbors helps improve the quality of feature representation; the decrease is small on Yelp Open with extremely sparse social relations, because the neighbors available for filtering are limited under sparse social networks. (2) Removing the social aggregation decreases Recall@10 by 0.4% on Yelp Open while remaining basically unchanged on LastFM-2k, indicating that social information aggregation contributes positively to user representation, and this contribution is more critical in sparse social scenarios. (3) Keeping only the interaction view decreases Recall@10 by 3.9% on Yelp Open but only 0.2% on LastFM-2k, demonstrating that multi-view item feature extraction is particularly important on Yelp Open with rich item attributes (business category entities). (4) Fixing the modality fusion weights to 0.5/0.5 decreases Recall@10 by 0.8% on Yelp Open, verifying the necessity of attention-driven dynamic fusion; on LastFM-2k, however, the performance slightly increases by 0.5%, possibly because the modality signals of this dataset are relatively balanced and fixed weights can achieve similar results, indicating that the advantage of dynamic fusion is more prominent in scenarios with significant modality differences and sparse interactions. (5) Removing the multi-layer information propagation decreases Recall@10 by 2.2% on LastFM-2k and 0.4% on Yelp Open, and decreases NDCG@10 by 3.6% on LastFM-2k and 3.2% on Yelp Open, showing that multi-layer propagation is crucial for modeling high-order correlations and improving ranking quality.

5. Conclusion

Aiming at the problems that traditional recommendation algorithms ignore the inconsistency of social neighbors and use only single-modal auxiliary information, this paper proposes a multi-view social recommendation algorithm fusing multimodal auxiliary information. Through user-consistent social modeling, the attention mechanism and a differentiable gate dynamically select preference-consistent social neighbors for neighborhood aggregation, alleviating the insufficient utilization of social information. Through multi-view item feature extraction, item representations are learned from both the knowledge graph view and the historical interaction view and integrated by a cross-attention fusion unit, enriching the semantic information of item features. Through multimodal dynamic fusion, attention-based late fusion of textual and visual modalities is performed on top of multi-layer information propagation, achieving scenario-adaptive modality weights. Experiments on two public datasets show that the proposed algorithm achieves the best performance on both datasets, with the advantage being particularly prominent on Yelp Open with sparse social relations, and ablation studies verify the effectiveness of each component. In future work, we will investigate finer-grained modality interaction mechanisms and validate the scalability of the algorithm on larger industrial datasets.

Acknowledgments

This research was funded by the Scientific Research Support Program of Sias University under grant numbers 2025XKD087.

References

[1] Resnick, P., & Varian, H. R. (1997). Recommender systems. Communications of the ACM, 40(3), 56-58. https://doi.org/10. 1145/ 245108.245126.

[2] Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE Transactions on Knowledge and Data Engineering, 17(6), 734-749. https://doi. org/ 10. 1109/TKDE.2005.99.

[3] Su, X. Y., & Khoshgoftaar, T. M. (2009). A survey of collaborative filtering techniques. Advances in Artificial Intelligence, 2009, 421425. https://doi.org/ 10. 1155/ 2009/ 421425.

[4] Wu, L., Sun, P. J., Hong, R. C., & Wang, M. (2018). SocialGCN: An efficient graph convolutional network based model for social recommendation. CoRR, abs/1811.02815.

[5] Yu, J. L., Yin, H. Z., Li, J. D., Wang, Q., & Hung, N. Q. V. (2021). Self-supervised multi-channel hypergraph convolutional network for social recommendation. In Proceedings of the Web Conference 2021 (pp. 413-424). https: // doi.org/10.1145/3442381.3449835.

[6] Baltrušaitis, T., Ahuja, C., & Morency, L.-P. (2019). Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2), 423-443. https://doi.org/10.1109/TPAMI.2018.2798607.

[7] Ma, H., Yang, H. X., Lyu, M. R., & King, I. (2008). SoRec: Social recommendation using probabilistic matrix factorization. In Proceedings of the 17th ACM Conference on Information and Knowledge Management (pp. 931-940). https:// doi.org/ 10. 1145/ 1458083.1458205.

[8] Yang, B., Lei, Y., Liu, J. M., & Li, X. (2017). Social collaborative filtering by trust. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(8), 1633-1647. https:// doi. org/10.1109/TPAMI.2016.2612259.

[9] Fan, W. Q., Ma, Y., Li, Q., He, Y., Zhao, E., Tang, J. L., & Yin, D. W. (2019). Graph neural networks for social recommendation. In Proceedings of the World Wide Web Conference (pp. 417-426). https://doi.org/ 10.1145/ 3308558. 3313488.

[10] Wu, L., Li, J. W., Sun, P. J., Hong, R. C., Ge, Y., & Wang, M. (2022). DiffNet++: A neural influence and interest diffusion network for social recommendation. IEEE Transactions on Knowledge and Data Engineering, 34(10), 4753-4766. https:// doi. org/10.1109/TKDE.2021.3092445.

[11] Zhang, F. Z., Yuan, N. J., Lian, D. F., Xie, X., & Sun, G. Z. (2016). Collaborative knowledge base embedding for recommender systems. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 353-362). https://doi.org/ 10. 1145/ 293 9672.2939673.

[12] Wang, X., He, X. N., Cao, Y. X., Liu, M., & Chua, T.-S. (2019). KGAT: Knowledge graph attention network for recommendation. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 950-958). https://doi.org/10. 1145/ 3292500. 3330989.

[13] He, R. N., & McAuley, J. (2016). VBPR: Visual Bayesian personalized ranking from implicit feedback. In Proceedings of the 30th AAAI Conference on Artificial Intelligence (pp. 144-150). https://doi.org/10.1609/aaai.v30i1.9894.

[14] Wei, Y. W., Wang, X., Nie, L. Q., & Chua, T.-S. (2019). MMGCN: Multi-modal graph convolution network for personalized recommendation of micro-video. In Proceedings of the 27th ACM International Conference on Multimedia (pp. 1437-1445). https://doi.org/10.1145/3343031.3350940.

[15] Rendle, S., Freudenthaler, C., Gantner, Z., & Schmidt-Thieme, L. (2012). BPR: Bayesian personalized ranking from implicit feedback. CoRR, abs/1205.2618.

[16] He, X. N., Liao, L. Z., Zhang, H. W., Nie, L. Q., Hu, X., & Chua, T.-S. (2017). Neural collaborative filtering. In Proceedings of the 26th International Conference on World Wide Web (pp. 173-182). https://doi.org/10. 1145/ 30389 12. 3052569.

[17] He, X. N., Deng, K., Wang, X., Li, Y., Zhang, Y. D., & Wang, M. (2020). 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 (pp. 639-648). https: // doi.org/10.1145/3397271.3401063.

[18] Jamali, M., & Ester, M. (2010). A matrix factorization technique with trust propagation for recommendation in social networks. In Proceedings of the 4th ACM Conference on Recommender Systems (pp. 135-142). https://doi.org/ 10. 1145/ 1864708.1864736.