A survey of multimodal recommendation systems

Anonymous CVPR submission

Paper ID ChuangHong Lin

Abstract

With the continuous development of network applications, the network resources are growing exponentially, and the phenomenon of information overload is becoming more and more serious. How to efficiently obtain the resources that meet the needs has become one of the problems that plague people. The recommendation system can effectively filter the massive information, and recommend the resources that meet their needs for the users.With the emergence of multimedia services such as short videos and news, it has become increasingly important to understand this content in recommendation. In addition, multimodal features also help alleviate the data sparsity problem in RS. Therefore, multimodal recommendation systems (MRS) have attracted wide attention from academia and industry in recent years.In this paper, we first introduced three traditional recommendation technologies, and then introduced the components of the MRS and the general process of MRS, and according to the different classification methods, introduced four multimodal recommendation systems. Finally, we discuss the challenges MRS Faces and summarize the paper.

1. Introduction

In recent years, the rapid development of network applications, especially mobile applications, makes it convenient for people to browse a large number of network information resources. How to recommend resources (such as commodities, movies, books, etc.) for users from massive information resources has become one of the concerns of researchers. Recommendation system (Recommendation System, RS) can effectively filter and filter information, help users to retrieve information resources that meet their needs in a personalized way, and alleviate the problem of information overload (Information Overload). After continuous development and update, recommendation technology has been widely used in education, music, e-commerce, social networking and other fields.

Due to the development of multimodal research [3], multimodal recommender systems (MRS) have been designed and applied in recent years. On the one hand, MRS can handle different modal information, which is inherent in multimedia services. On the other hand, MRS can also utilize rich item multimodal information to alleviate the data sparsity and cold start problems that are widely present in recommender systems.

2. Traditional recommendation algorithm

Recommendation system is a new research field combined by data mining, prediction algorithm [2], machine learning and other disciplines. Literature[6]in the earliest definition of recommendation system, points out that in daily life whether understand events or unknown events, always need people to make decisions, in the face of familiar things, people can often rely on past experience to make reasonable decisions, however, in the face of the unknown things, people need others oral advice, book reviews, reviews, recommendation, etc, the literature that the significance of the recommendation system is able to recommend project and users to establish appropriate matching relationship. In literature [16], recommendation system is a project that matches different users from a large number of projects to users that match their preferences but are not observed by users. It believes that recommendation system is becoming an important business with significant economic impact.

In essence, the recommendation system is a simulation of a certain human behavior. It analyzes and processes the specific data information through the recommendation algorithm, and then recommends the processed results to the user [9] with relevant needs. Recommendation algorithm is the core of the recommendation system. It can model the preferences according to users 'historical purchase needs, behavior records or similarities, so as to find the needs that meet users' preferences and recommend them to users. The formal definition of the recommendation system [1]is as follows.

2.1. Recommendation technology based on content filtering

The recommendation system was first applied in e-commerce websites. It usually recommends items [10] to users with similar demand preferences based on their purchase behavior records or purchase evaluation. A context-based approach to matching and sort services is proposed in literature [15], arguing that context is the relevant set of linguistic terms used to describe a given text. The method extracts the tokens as text terms by parsing the underlying documents and uses a string matching function to match the ontology of these tokens. A service discovery method that matches the user query and service description and relevant contextual information is proposed in literature [6]. This method models the context information provided by the context provider, the service description provided by the service provider, and the service request provided by the user with the ontology, and then matches the three information one by one. A Web service context classification is proposed in literature [13], and then an ontology is used to define this classification. Context is modeled by a two-level mechanism that covers the context specification and service strategy, providing a peer-to-peer architecture to fully match the Web service context strategy, and each context of the source service is matched by the strategy of the candidate service.

In short, the core idea of the recommendation (CB) technology based on content filtering is to take the selection record or preference record of the user history as the reference recommendation, and to mine the items with high correlation with the reference recommendation in other unknown records as the content of the system recommendation. The interaction records of users in a certain period of time are obtained through explicit feedback (e. g., evaluation, approval, liking), browsing time, clicks, search time, stay time, etc.), then learn the preferences of the users in these records and mark the characteristics of the content (or matching degree); finally ranking the similarity between the recommended objects to be tested and the user preferences, thus selecting the recommendations according to their preferences. Calculating similarity is a key part that directly affects the recommended strategy. There are many ways to calculate similarity, common formula (2) calculate similarity:

$$u(p,c) = score(userprofile, content) \tag{1}$$

Where: p represents the user, c represents the recommended content, userprofile indicates the preferred content, content represents the content recommended by the user.score It is used to calculate the similar values of user preferences and recommended content, and it is finally defined by the utility function u (). According to the value of u, the larger the value is, the higher the ranking is.

The calculated u value is sorted, and the larger the u value is, the more the recommended object conforms to the user's preference. For example, when recommending movies for users, the system will learn the user's historical viewing records and analyze them, then find the commonalities of these movies, predict the type of movie that the user is interested in, and then select movies similar to the user's preferences from the massive movie list. The characteristic marking and recommended content of user preference records are the key of CB, and user evaluation has less influence on content-based recommendation system.

2.2. Collaborative filtering recommendation

The core of the collaborative filtering recommendation (CF) algorithm is to obtain the dependency relationship between users and projects by analyzing the scoring matrix (usually the score of users on the project), and further predict the correlation relationship between the new user and the project. The CF algorithm is one of the first recommendation techniques to be studied and discussed, and it effectively promotes the development of personalized recommendation. In 1992, document [7] used traditional collaborative filtering technology to solve the spam classification problem; Amazon (Amazon) is one of the largest online shopping platforms, mainly using CF algorithm to recommend products to users; Netflix also uses CF algorithm to recommend their favorite TV programs on its home page. Nowadays, collaborative filtering technology is widely used in music recommendation, film recommendation, e-commerce and other fields [4]. CF is mainly divided into memory-based (Memory-Based) recommendation and model-based (Model-Base-based) recommendation.

2.2.1 Memory-based recommendations

Memory-based collaborative filtering recommends finding the similarity between similar users and similar items [6, 8] through the evaluation matrix of user-item (User-Item), and then building a similarity matrix for new users to predict the items of interest of users. Recommendation by finding similar items is called project-based recommendation; recommendation by finding similar users is called user-based recommendation. The project-based collaborative filtering technology mainly excavates and analyzes the hidden relationship between different recommended projects rather than the relationship between users [11]. The similarity calculation between projects is the key of this technology, and the recommendation process is shown in Figure 1. The process can be understood as follows: if there are two different users A and B, and they all show a high love for item 1 and 3, then we can think that there is some similarity between item 1 and 3. When a new user C appears in the system and selects item 1, the system will automatically recommend item 3 with high similarity to item 1 to him.

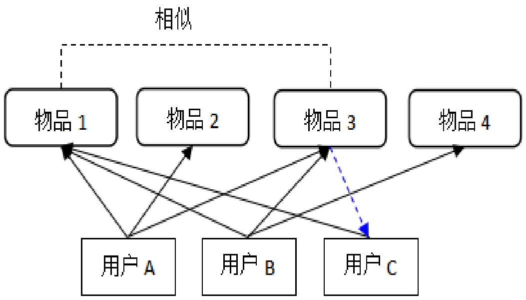

Figure 1. Item-based collaborative filtering recommendation

Based on the user recommendation process as shown in figure 2, after the evaluation matrix calculation, think user A is similar to B, when item selection, if the user A selected item 1,2,3, user B selected item 1,3, then in the item recommendation can think user B selection and user A similar, so the recommendation system can recommend item 2 to the user B.

In literature [14], the user matrix is analyzed to determine the differences between these users and users and different users and the projects they are interested in, so as to recommend appropriate projects for users according to the differences. However, the recommendation process based on users cannot rely on similar users to know each other. Therefore, literature [16]proposes a collaborative filtering algorithm based on anonymous cooperation, which is specifically used to solve the problem of recommending news and movies for different users. Although the user-based collaborative filtering algorithm can find the hidden interests and preferences of users, the technology has serious cold start problems. In practical problems, the type of users in the recommendation system is not invariable. When a new user type appears, the system lacks the user's preference record, so the recommendation system cannot provide the users with recommendations that meet their needs.In order to solve the cold start problem faced by collaborative filtering, the traditional collaborative filtering algorithm and neural network algorithm are combined in literature [17]. Neural network algorithm is one kind of deep learning algorithm, which can analyze and calculate the complex nonlinear relationship between users and the project, with high efficiency. The mixed model in literature [17]focuses on the typicality and diversity of the recommended objects. After evaluation in the application of the Korean national health and nutrition survey data, the results show that it can indeed improve the recommendation effect.

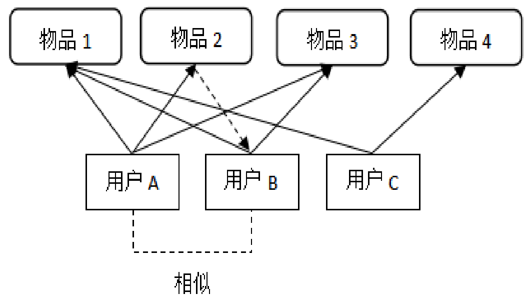

Figure 2. User-based collaborative filtering recommendation

2.2.2 Model-based recommendations

The model-based recommendation algorithm is to predict the user's score of uninteracting items by training mathematical models, usually including probability matrix decomposition (Probabilistic Matrix Factorization, PMF) [13] and singular value decomposition (Singular Value Decomposition, SVD). The main idea of PMF and SVD is to establish an appropriate model for the historical interaction data record between the user and the project, and then produce a list of recommendations that meet the needs of the user, among which the recommendation based on matrix decomposition is widely used. The PMF model generally believes that the interaction behavior of the user and the recommended item is only determined by a few factors potentially affecting their interest preferences. Therefore, the higher-order scoring matrix Rnm is decomposed into two low-dimensional matrices E and Q, as shown in Equation (2):

$$R \approx E^T Q \tag{2}$$

Where: E = (e1, e2,..., en) represents the low-dimensional user feature matrix, ei represents the k-dimensional feature vector of the user i; Q = (q1, q2,..., qn) represents the low-dimensional recommended item feature matrix.

2.3. Mixed recommendation

Content-based recommendation technology often reduces large-scale information content over time; collaborative filtering technology is easy to encounter cold start problems in new projects; and hybrid recommendation technology is a recommendation method to avoid different advantages and disadvantages, and integrates different algorithms into the recommendation system, that is, mixed recommendation [5, 12]. The current hybrid recommendations are mainly divided into pre-fusion, post-fusion, and medium fusion.

3. Multi-mode recommendation system

Multimodal recommender system refers to a type of recommender system that utilizes multiple sources of data (i.e., multimodal data) for recommendation. These sources of data include various forms of data such as text, image, audio, and video. Compared to traditional single-modal recommender systems, multimodal recommender systems can better understand users' needs and interests comprehensively, thereby providing more accurate and personalized recommendation services.

3.1. Multimodal recommendation system component

Data acquisition: Multi-modal data is collected from different data sources, such as users' browsing history, purchase records, social media data, etc.

Data fusion: Data from different sources are fused to form multi-modal data sets. Feature extraction: Feature extraction is carried out on multi-modal data for subsequent recommendation calculation. For example, convolutional neural network (CNN) can be used for feature extraction of image data, and cyclic neural network (RNN) can be used for feature extraction of text data.

Recommendation computing: Machine learning algorithms and recommendation algorithms are used to analyze and process multi-modal data in order to generate personalized recommendation results.

Recommendation display: Display the recommendation results to users and collect feedback data from users to continuously optimize the performance of the recommendation system. The multi-modal recommendation system can be applied in many fields, such as e-commerce, social media, online advertising, etc., to provide users with more personalized and comprehensive recommendation services.

3.2. General flow of multimodal recommendation system

Feature extraction: In multi-modal recommendation, each item to be recommended includes two types of features. One is tabular features, such as the id of the item, category, and so on. The other is multimodal features, including descriptive pictures of items, evaluation text, etc. At this stage, multimodal recommendation systems use modal encoders to encode multimodal features, such as Vits for picture processing and Bert for text processing.

Feature interaction: The representation vectors of different modal features obtained from feature extraction are usually in different semantic Spaces, and users have different preferences for different modes. Therefore, in this stage, the multi-modal recommendation system interacts and integrates the multi-modal representation to obtain the representation vector of items and users.

Recommendation: After obtaining the representation vector of users and items, the recommendation model can be used to calculate the recommendation probability and output the recommendation list.

Taking fitness apps as an example, this application combines various modalities of information such as text, images, and videos to assist users in better fitness training. Specifically, the application can establish user profiles by analyzing body data, fitness goals, and exercise habits. Additionally, it can create exercise profiles by analyzing textual descriptions, images, and videos of exercises. Then, through multimodal fusion technology, the system matches user profiles with exercise profiles to generate personalized fitness plans. For instance, if a user wants to lose weight, the system can recommend appropriate exercises for weight loss and suggest suitable fitness plans based on the user's physical condition and exercise habits to meet their needs. This example demonstrates the application of multimodal recommendation systems in the fitness field, which can help users improve their physical health and fitness training.

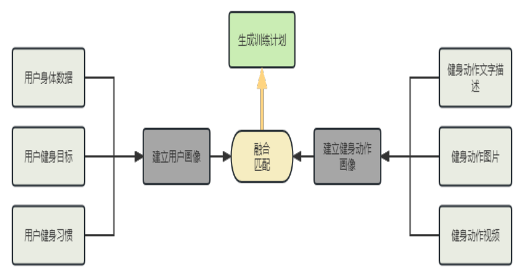

Figure 3. Fitness application recommendation process

4. Cslasification

According to the different ways, we give the following multimodal recommendation system classification.

4.1. Sort by data type

According to different data types, multimodal recommendation systems can be divided into different types such as text-image, image-image, and text-audio. Text-image multimodal recommendation systems are mainly applied in e-commerce and social fields, providing users with more accurate product recommendations by combining product text descriptions and images. Image-image multimodal recommendation systems are mainly applied in tourism and food fields, providing users with more personalized recommendations by combining user-uploaded images and relevant image databases. Text-audio multimodal recommendation systems are mainly applied in the music field, providing users with more accurate music recommendations by combining song text descriptions and the songs themselves.

4.2. Sort by fusion mode

According to different fusion methods, multimodal recommendation systems can be divided into different types such as feature fusion, decision fusion, and hybrid fusion. Feature fusion refers to the fusion of feature vectors from different modalities to obtain a comprehensive feature vector, which is then used for recommendation through machine learning models. Decision fusion refers to the fusion of recommendation results from different modalities to obtain a comprehensive recommendation result. Hybrid fusion refers to the use of both feature fusion and decision fusion methods for recommendation.

4.3. Sort by application domain

According to different application areas, multimodal recommendation systems can be divided into different types such as e-commerce recommendation, social recommendation, fitness recommendation, and travel recommendation. E-commerce recommendation systems provide users with more accurate product recommendations by combining product text descriptions and images. Social recommendation systems provide users with more personalized recommendations by combining user social relationships and interests. Fitness recommendation systems provide users with more personalized fitness training plans by combining user body data and fitness goals. Travel recommendation systems provide users with more personalized travel route recommendations by combining user travel time, destination, and preferences.

4.4. Sort by recommended target

According to different recommendation targets, multimodal recommendation systems can be divided into different types such as product recommendation, user recommendation, and advertising recommendation. Product recommendation systems refer to recommending products to users to increase sales and user satisfaction. User recommendation systems refer to recommending users to other users or social networks to increase user activity and social effects. Advertising recommendation systems refer to recommending advertisements to users to improve advertising effectiveness and return on investment.

5. Key technology research

5.1. Feature interaction

Multi-modal data refers to various modalities that describe information. Because they are sparse and have different semantic spaces, connecting them to recommendation tasks is essential. Feature interaction can transform different feature spaces into a unified semantic space through non-linear transformation, ultimately improving the performance and generalization ability of recommendations.

5.1.1 Merge

In multi-modal recommendation scenarios, there is a large number of multi-modal information types and quantities for users and items. Therefore, it is necessary to integrate different multi-modal information to generate feature vectors to serve recommendation models. Compared with bridging, fusion pays more attention to the multi-modal relationships within items. Specifically, it aims to integrate various preferences and patterns. Since the inter-item and intra-item modal relationships are crucial for learning item representations, many MRS models even adopt both fusion and bridging. Attention mechanism is the most widely used feature fusion method, which can flexibly combine multi-modal information according to attention and interest.

5.1.2 Bridging

Bridging here refers to the construction of multi-modal information transmission channels. It focuses on capturing the interaction between users and items based on multi-modal information. The difference between multi-modal recommendation and traditional recommendation is that items contain rich multimedia information. Early research simply used multi-modal content to enhance item representation, but they often ignored the association between users. Graph neural networks can capture the interaction between users and items through message passing mechanisms, thereby enhancing user representation and further capturing user preferences for different modal information.

5.2. Multimodal feature enhancement

Different modal representations of the same object have unique and common semantic information. If these two features can be distinguished, the recommendation performance and generalization ability of MRS can be significantly improved. Recently, to address this issue, some works have proposed Disentangled Representation Learning (DRL) and Contrastive Learning (CL) for interaction-based feature enhancement.

6. Challenge

Data fusion: Different modalities of data need to be fused, but different fusion methods and their effects need to be considered. For example, feature fusion, decision fusion, and hybrid fusion methods all have their own advantages and disadvantages, and the appropriate fusion method should be selected based on the specific situation.

Data quality: Data quality varies among different modalities, and it is a challenge to handle data of different qualities and prevent noise from affecting the recommendation effect. For example, data cleaning and preprocessing can be used to effectively improve data quality and reduce the impact of noise.

User preferences: User preferences are often multi-modal, and it is a challenge to effectively fuse multiple modalities of preferences and accurately reflect user preferences. For example, weighted averaging or model-based methods can be used to fuse user preferences.

Model selection: Different models are suitable for different data types and fusion methods, and it is a challenge to choose the appropriate model. For example, deep learning models or traditional machine learning models can be chosen based on the different data types and fusion methods.

Real-time performance: Multimodal recommendation systems need to respond to user requests in real-time, and it is a challenge to ensure real-time performance while maintaining recommendation effectiveness. For example, caching, preprocessing, and distributed computing can be used to improve system response speed.

Privacy protection: Multimodal recommendation systems need to handle various types of user data, and it is a challenge to protect user privacy. For example, data encryption, differential privacy, and model distillation can be used to protect user privacy.

7. Conclusion

Multimodal recommendation systems, with their aggregation advantages across different modalities, are becoming one of the forefront research directions in recommendation systems.

References

[1] Gediminas Adomavicius and Alexander Tuzhilin. 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, 2005. 1

[2] Jon Scott Armstrong. *Principles of forecasting: a handbook for researchers and practitioners*, volume 30. Springer, 2001. 1

[3] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018. 1

[4] Yi Cai, Ho-fung Leung, Qing Li, Huaqing Min, Jie Tang, and Juanzi Li. Typicality-based collaborative filtering recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 26(3):766–779, 2013. 2

[5] Luis M De Campos, Juan M Fernández-Luna, Juan F Huete, and Miguel A Rueda-Morales. Combining content-based and collaborative recommendations: A hybrid approach based on bayesian networks. *International journal of approximate reasoning*, 51(7):785–799, 2010. 3

[6] Mukund Deshpande and George Karypis. Item-based top-n recommendation algorithms. *ACM Transactions on Information Systems (TOIS)*, 22(1):143–177, 2004. 2

[7] David Goldberg, David Nichols, Brian M Oki, and Douglas Terry. Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, 35(12):61–70, 1992. 2

[8] Yehuda Koren. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 426–434, 2008. 2

[9] Hongyan Liu, Jun He, Tingting Wang, Wenting Song, and Xiaoyang Du. Combining user preferences and user opinions for accurate recommendation. *Electronic Commerce Research and Applications*, 12(1):14–23, 2013. 1

[10] Liwei Liu, Freddy Lecue, and Nikolay Mehandjiev. Semantic content-based recommendation of software services using context. *ACM Transactions on the Web (TWEB)*, 7(3):1–20, 2013. 2

[11] Wenming Ma, Junfeng Shi, and Ruidong Zhao. Normalizing item-based collaborative filter using context-aware scaled baseline predictor. *Mathematical Problems in Engineering*, 2017, 2017. 2

[12] Michael J Pazzani. A framework for collaborative, content-based and demographic filtering. *Artificial intelligence review*, 13:393–408, 1999. 3

[13] JC Platt, D Koller, Y Singer, and ST Roweis. Proceedings of the 20th international conference on neural information processing systems, 2007. 2, 3

[14] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*, pages 285–295, 2001. 3

[15] Aviv Segev and Eran Toch. Context-based matching and ranking of web services for composition. *IEEE Transactions on Services Computing*, 2(3):210–222, 2009. 2

[16] Mingxuan Sun, Guy Lebanon, and Paul Kidwell. Estimating probabilities in recommendation systems. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 734–742. JMLR Workshop and Conference Proceedings, 2011. 1, 3

[17] Hyun Yoo and Kyungyong Chung. Deep learning-based evolutionary recommendation model for heterogeneous big data integration. *KSII Transactions on Internet and Information Systems (TIIS)*, 14(9):3730–3744, 2020. 3