ISSN: 2688-8645

The Sequence Recommendation Algorithm based on GRU and Attention

Bo He[a], Kaiwei Zhu[b, *], Qingyang Lai[c]

School of Computer Science and Engineering, Chongqing University of Technology, Chongqing, 400054, China

[a]hebo@cqut.edu.cn, [b, *]zkw3019cg@163.com, [c]715078797@qq.com

Abstract

In recent years, with the booming development of social networks and e-commerce, users have increasingly convenient access to information, while a large amount of data continues to emerge, leading to more and more serious information overload. To alleviate the problem of information overload, recommendation systems have emerged to assist users in sifting through massive amounts of information to find content that meets their needs. Sequential recommendation, as a form of recommendation system, mainly analyzes the interaction behavior between users and items, models user characteristics, and then uses various methods to capture users' long-term and short-term preferences to recommend items of interest to users. Based on the perspective of user preference change over time, this paper provides an in-depth analysis of the current research progress and methods of user behavior sequence recommendation. At the same time, this paper proposes corresponding solution strategies for the problems of cold start, sparse matrix and noise interference faced by traditional recommendation systems. Finally, we will discuss the challenges and future research directions of recommendation systems to provide the theoretical basis for further improvement of recommendation systems.

Keywords

Sequential Recommendation; Long-term Preference; Short-term Preference; Cold Start.

1. Introduction

With the rapid development of the Internet and big data, the emergence of massive amounts of data has created a huge challenge for users in meeting their needs, resulting in information overload. Although the traditional search engine technology can solve the user's information retrieval needs to a certain extent, it can only search by keywords and can only return the same search results for the same search keywords, which is difficult to meet the diverse needs of users. In contrast, recommendation systems can effectively filter and filter information, provide users with information resources that meet their needs in a personalized manner, and alleviate the problem of information overload to a certain extent. Personalized recommendation technology helps users discover their potential needs from a large amount of information by studying their interests and preferences. Collaborative filtering recommendation algorithm has developed rapidly since it was proposed in 1992 and has received a lot of attention from academics, especially the introduction of collaborative filtering algorithm in personalized recommendation technology has become a hot research trend in recent years. This algorithm has been widely used in advertising, movie recommendation and e-commerce. The advantages of recommendation systems are to improve user experience and to explore the "long tail" of commodities and serve the economy and society. However, recommendation systems need to collect user information as detailed as possible to build user models in order to provide

ISSN: 2688-8645

personalized recommendation services. Although collaborative filtering has been used as a typical recommendation technique, it still faces many problems that need to be solved.

Sequential recommendation is a recommendation algorithm that mainly models user behavior sequences, which focuses more on capturing users' transient and instantaneous preferences. By using the temporal information of user interactions, sequential recommendation can effectively capture user preferences and achieve better recommendation results, and many things exist in real life in the form of sequences[1] For example, when users used to prefer foreign branded cell phones, but with the development of domestic branded cell phones, users pay more attention to domestic branded cell phones, so users browse more domestic branded cell phones. In contrast, traditional recommendation algorithms have difficulty in quickly capturing changes in user interests, while sequential recommendations are able to capture the dynamic preferences of users through their interactions with items. Traditional sequential recommendation algorithms, such as collaborative filtering algorithms, focus on modeling explicit or implicit interactions between users and items to obtain users' preferences. The algorithm tends to use the user's historical interactions to predict the user's static preferences, but the user's preferences are dynamic and there may be dependencies on the information of the before and after interactions in the sequence. Thus, the user's next behavior depends not only on their static long-term preferences, but also largely on their short-term preferences. To better capture the dynamic user preferences, modern sequential recommendation algorithms consider not only the interaction between users and items, but also the temporal information of the interaction, such as the order and time interval in the user behavior sequence. These algorithms can predict user preferences more accurately and provide more personalized recommendation services to users.

This paper introduces various existing common sequence recommendation algorithms, and classifies sequence recommendation into: user short-term preference sequence recommendation, user long-term preference sequence recommendation, and user long-term and short-term preference sequence recommendation. The main problems faced in this field are summarized.

2. User Long-term Preference Sequence Recommendations

2.1. Collaborative Filtering Recommendations

The collaborative filtering recommendation algorithm is one of the widely used recommendation algorithms in major recommendation platforms, and has had a profound impact on the development of recommendation systems since its introduction. The core idea is to find similar nearest neighbors based on the target user's interest preferences, so as to make recommendations. Among collaborative filtering algorithms, matrix decomposition (MF) based recommendation algorithm is the core of it. This technique constructs a scoring matrix of user and item features and reconstructs the matrix by inner product. However, it is difficult to make more accurate recommendations because deeper correlations cannot be uncovered by relying only on the basic user and item features.

Collaborative filtering recommendation algorithms can be divided into two types of algorithms, user-based and item-based[2]. The basic idea of the user-based collaborative filtering algorithm is to establish the preference relationship between users and items by mining massive user historical behavior data, and find the neighboring user groups with similar historical interest preferences to the target users. By using the historical preference data of the neighboring user groups, the target user's rating of the items preferred by the neighboring users is estimated, and then the highest rated items are recommended to the target user. The main idea of this algorithm is to first identify the target users, and find users with similar interest preferences to them, and recommend the content that these similar users are interested in to the target users.

ISSN: 2688-8645

This algorithm is suitable for applications with few users, such as news recommendations. The item-based collaborative filtering algorithm mainly analyzes users' historical behavior records to find items similar to them and recommends this category of items to users. This algorithm is suitable for applications where the number of items is much smaller than the number of users, such as books and e-commerce. The collaborative filtering classification algorithm is shown in Figure 1.

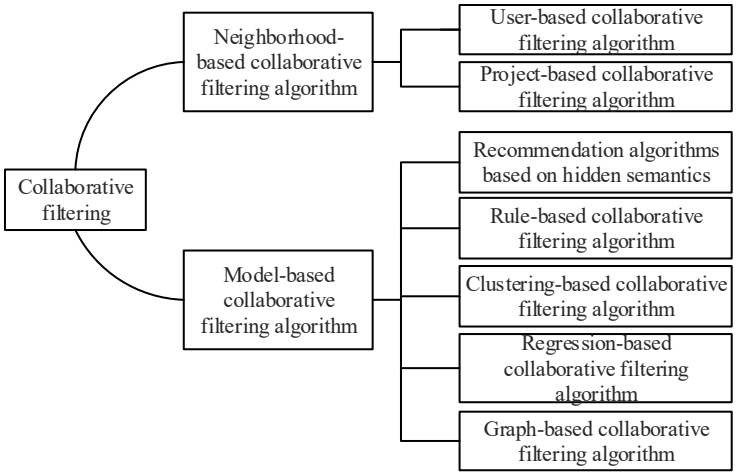

Figure 1. Schematic diagram of collaborative filtering algorithm classification

2.2. Recurrent Neural Network Sequence Recommendation

Recurrent neural networks can improve the accuracy of recommendation systems by modeling the sequential data of user behavior to learn the dynamic interest changes of users. A large number of sequential recommendation methods are based on recurrent neural networks. The standard recurrent neural network has the advantages of low number of parameters and simple structure. However, this module design has huge drawbacks. As the length of a time series increases, recurrent neural networks quickly forget the early input states[3] For complete time series, it is difficult for recurrent neural networks to learn the potential relationships between large span and long-range sequence positions because of its "forgetfulness" property. Time series data are characterized by long-term dependence, but recurrent neural networks cannot solve this problem. The literature[4]uses recurrent neural networks to learn users' long-term interests from their historical behavior sequences, which has important implications for personalized recommendations.

Long short-term memory (LSTM) is a variant of RNN that can learn long-term dependencies. Through the selection of forgetting and input gates, important information can be maintained and transmitted in the network. However, the LSTM model is more complex and has more parameters. GRU is a variant of recurrent neural network, which is more complex than standard recurrent neural network but simpler than long short-term memory network, and more suitable for modeling processed collaborative filtering data in terms of accuracy and efficiency.GRU retains the advantages of long short-term memory network in learning long-term dependence, but has a simpler structure and is easier to train.

In the literature[5], a sequence recommendation model based on gated recurrent units (GRU For Recommendation, GRU4Rec) is proposed to apply recurrent neural networks to sequence recommendation to model user behavior sequences. The model uses a single gate to simultaneously control the forgetting factor and update the state unit for predicting the next user behavior in a session. Song et al[6] proposed an enhanced recurrent neural network model to enhance the existing recurrent neural network sequence recommendation model by

ISSN: 2688-8645

extracting higher-order user contextual preferences.Kumar et al[7] proposed a Recurrent Attention Deep Semantic Structured Model. RADSSM), which uses a bidirectional long and short-term memory network to efficiently obtain the information contained in a sequence.

3. User Short Term Preference Sequence Recommendation

3.1. Markov Chain Sequence Recommendation

A Markov chain-based sequential recommendation model generates a sequential model by modeling a sequence of user-item interactions and predicts the next potential item to interact with. In a first-order Markov chain, the value of the observed variable v is influenced only by its associated state variable V, while the behavior at moment t depends only on the behavior at moment $t-1$ and is independent of the previous $t-2$ behaviors. In a recommender system, the next interaction depends only on the current interaction and is independent of historical interactions.Rendle et al[8] applied Markov chains to a short sequence recommendation model that captures transitions between items and performs well with a certain degree of data sparsity.Hosseinzadeh et al[9] modeled user behavior as a hidden Markov chain (Hidden Markov Model (HMM) to capture changes in user preferences and model the current contextual information of the user as a hidden variable in the model. He et al[10] used a higher-order Markov chain model and considered the last few user interactions. Although higher-order Markov chains consider more historical interaction dependencies, it still cannot capture more behavioral dependencies in relatively long sequences and cannot reflect users' long-term preferences. In addition, it also performs poorly in handling sparse data.The interactions between multiple items in realistic sequences are its important features.

3.2. Convolutional Neural Network Sequence Recommendation

Convolutional Neural Network (CNN) is a class of neural network architecture commonly used in deep learning. CNNs are composed of an input layer, a convolutional layer, a pooling layer, a fully connected layer and an output layer. The convolutional and pooling layers cooperate to achieve feature extraction. In the convolutional layer, the neurons in the upper layer connect only some of the neurons in the next layer. In addition, CNN uses weight sharing, i.e., convolutional kernel, which not only reduces the number of parameters significantly, but also reduces the risk of overfitting. The local connections and weight sharing in the convolutional layer make the number of parameters of the CNN model better than that of the fully connected layer network, easier to train, and harder to fit. Therefore, CNN models have achieved good results in the fields of computer vision, natural language processing and recommendation systems.

Tang et al[11] proposed a Convolutional Sequence Embedding Recommendation Model (Caser), which extracts several consecutive terms from a sequence as input to embed them into a neural network, captures local features of the sequence using horizontal and vertical convolutional layers, and then obtains higher-level features through fully connected layers. The model extracts several consecutive terms from the sequence as input and embeds them into the neural network. The convolutional neural network extracts the features of the current sequence, and instead of processing the data directly from left-to-right one-way structure, the model gives the whole as structured information to the CNN for feature extraction. Yuan et al[12] improved the Caser model, which is a shallow network that has difficulty capturing complex relationships as well as modeling long-term dependencies, and considers only the conditional probability distribution of the last item when generating the next item.

Table 1 summarizes the main studies in the user short-term preference sequence recommendations in this paper.

ISSN: 2688-8645

Table 1. Main studies of sequential recommendation of users' short-term preferences

Category	Method	For questions	Main advantages
Markov chain sequence recommendation	Rendle et al.[8]	Tidligere anbefalingssystemer har laget prediksjoner basert på historiske datasett.	Den første anvendelsen av Markov-kjeder på anbefalinger, som tar hensyn til den nåværende interaksjonen
	Hosseinzadeh et al.[9]	Det er ikke bare sekvensen i seg selv som påvirker anbefalingens effektivitet.	Forutsi brukerpreferanser med hensyn til dynamiske, partiske brukere og partiske kontekster
	He et al.[10]	Vanskeligheter med å håndtere komplekse bruker-objekt-interaksjoner	Håndterer store og sparsomme datasett godt
Convolutional Neural Networks Sequence Recommendation	Caser.[11]	De fleste anbefalingssystemer anbefaler varer basert på generelle brukerpreferanser.	Fleksibel nettverksstruktur for å fange opp viktige trekk i sekvensen
	Yuan et al.[12]	Sekvensielle anbefalingsmodeller basert pa tilbakevendende nevrale nettverk baserer seg ofte på skjulte tilstander i fortiden og har begrenset treningshastighet.	Avbøter problemet med gradienter som forsvinner

4. User Long and Short Term Preference Sequence Recommendation

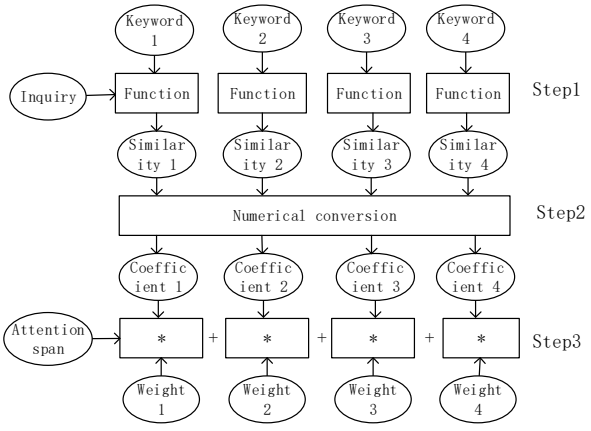

Figure 2. Self-attention calculation chart

In 2017, Google researchers Vaswani et al. proposed the Transformer model and applied it to machine translation. the Transformer model introduces a self-attentive mechanism, which is a new attention mechanism that is better at finding intrinsic associations between input sequences The Transformer model introduces the self-attentive mechanism, which is a new

ISSN: 2688-8645

attention mechanism that is better at finding the intrinsic association between the input sequences without relying on external information. For example, a word vector q is input into the model, and the output vector p is obtained by the self-attentive mechanism. p is obtained by calculating the correlation between all the input vectors and q and then weighting them. The self-attentiveness calculation is shown in the figure 2.

Liu et al[13] designed the ShortTerm AttentionMemory Priority (STAMP) model, which proposes a new attention mechanism that can calculate attention weights based on the context of the session sequence and capture both long-term and short-term preferences of the user.The STAMP model requires separately Obtain the user's long-term preference and short-term preference, m_s is the average value of the user's click sequence, and m_t is the user's most recent click, indicating the user's current interest, as in Equation (1) and Equation (2).

$$m_s = \frac{1}{t}\sum_{i=1}^{t} x_i \tag{1}$$

$$m_t = x_t \tag{2}$$

The user's long-term preference is obtained by a simple Feedforward Neural Network (FNN), inputting m_s and m_t into the FNN to get the attention of each item, and the attention coefficient of item x_i is calculated by the formula, so that we can calculate

calculate the user's long-term preference m_α, as Eq:

$$\alpha_i = W_0\sigma(W_1 x_i + W_2 m_t + W_3 m_s + b_a) \tag{3}$$

$$m_a = \sum_{i=1}^{t} \alpha_i x_i \tag{4}$$

where W_0 is the weight vector, W_1, W_2, W_3 are the weight matrices, b_a is the bias vector, and $\sigma(\cdot)$ is the sigmoid function. The m_a and m_t are input into Multilayer Perceptron (MLP) for processing to extract features, respectively, and the obtained hidden representation vectors are h_s,and h_t, respectively, and for each candidate item, the score formula is calculated as

$$\hat{z}_i = \sigma(< h_s, h_t, x_i >) \tag{5}$$

In which, 3 vectors h_s, h_t and $x_i$are used for dot product operation, and $\hat{z}_i$ represents the non-normalized cosine similarity, the similarity of item x_i with the user's general interest and current interest, and finally the output is obtained by normalizing it using the softmax function.

5. Problems in Sequence Recommendation Studies

5.1. Cold Start Problems

The cold start problems of recommendation algorithms can be broadly classified into the following 3 categories:

(1) User cold start: i.e., how to recommend items to new users entering the system, when a new user first enters the recommendation system, the system does not have his historical preferences and behavioral information;

(2) Item cold start: i.e., dealing with how to recommend new items added to the recommendation system to users who may be interested in them;

(3) System cold start: i.e., how to develop a personalized recommendation system on a newly established website, when the number of users and user preferences on the website are small and only some information about the products is available[14].

5.2. Noise Information Interference Problem

Sequences themselves are rich in information, and the amount of information is proportional to the length of the sequence. However, there is a large amount of interaction-independent noisy information in the sequence, which interferes with the process of capturing user preferences. Current research shows that each item in the sequence has a dependency relationship with its neighboring items, and this dependency relationship can affect the

ISSN: 2688-8645

recommendation effect. However, when modeling such dependencies, the noise may lead to errors in modeling the dependencies. In addition, there are usually less relevant items in the sequence, and these items can further trigger false dependencies [15]. To address these challenges, some scholars have proposed an approach based on a two-layer attention network. The method aims to assign different importance weights to different items in the sequence, emphasize items with high relevance, and reduce the influence of noise on the model. As a result, the method can significantly improve the accuracy of the recommendation algorithm.

6. Conclusion

User behavior sequences show sequential nature in the time dimension, while their preferences change over time. This paper focuses on the advantages and disadvantages of various recommendation algorithms in implementing recommendation process and their applications in different scenarios. The article analyzes the sequential recommendation problem in three aspects: long-term sequential recommendation, short-term sequential recommendation, and long- and short-term sequential recommendation, and describes the main problems in this field of research and the model optimization methods for different scenarios. This paper tries to give an outlook on the future development trend of sequence recommendation. Contextual factors such as time, climate, geographic location, season, moment of the day, and scenario may have a significant impact on the user's final choice. By extracting information from the context, we can better understand users' needs and interests, and thus help to provide them with accurate recommendations. On the other hand, there is also a close relationship between different domain items. It is possible to consider using the source domain data to generate the desired recommendations in the target domain and extracting features belonging to different domains from the sequence data. As the research progresses and the application expands, sequence recommendation will face more demands and challenges. The development of recommendation systems is closely related to the challenges faced. In addition, combining techniques from outside the domain with sequence recommendation to solve the current sequence recommendation problem is a valuable research direction. In summary, sequence recommendation research is still a hot research topic in the field of intelligent information processing such as data mining, information retrieval, and machine learning.

Acknowledgments

This research is supported by the humanities and social science research project of Chongqing municipal education commission under grant No. 23SKGH252 and the Chongqing university of action plan for high quality development of postgraduate education of Chongqing university of technology under grant No. gzlcx20233259.

References

[1] Wang S, Hu L, Wang Y, et al. Sequential recommender systems: challenges, progress and prospects [J]. arXiv preprint arXiv:2001.04830, 2019.

[2] Sarwar B, Karypis G, Konstan J, et al. Item-based collaborative filtering recommendation algorithms[C]//Proceedings of the 10th international conference on World Wide Web. 2001: 285-295.

[3] Donkers T, Loepp B, Ziegler J. Sequential user-based recurrent neural network recommendations [C]//Proceedings of the eleventh ACM conference on recommender systems. 2017: 152-160.

[4] Li Z, Zhao H, Liu Q, et al. Learning from history and present: Next-item recommendation via discriminatively exploiting user behaviors[C]//Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining. 2018: 1734-1743.

[5] Hidasi B, Karatzoglou A, Baltrunas L, et al. Session-based recommendations with recurrent neural networks[J]. arXiv preprint arXiv:1511.06939, 2015.

[6] Song Y, Lee J G. Augmenting recurrent neural networks with high-order user-contextual preference for session-based recommendation[J]. arXiv preprint arXiv:1805.02983, 2018.

[7] Kumar V, Khattar D, Gupta S, et al. Deep Neural Architecture for News Recommendation[C]//CLEF (Working Notes). 2017.

[8] Rendle S, Freudenthaler C, Schmidt-Thieme L. Factorizing personalized markov chains for next-basket recommendation[C]//Proceedings of the 19th international conference on World wide web. 2010: 811-820.

[9] Hosseinzadeh Aghdam M, Hariri N, Mobasher B, et al. Adapting recommendations to contextual changes using hierarchical hidden markov models[C]//Proceedings of the 9th ACM Conference on Recommender Systems. 2015: 241-244.

[10] He R, Fang C, Wang Z, et al. Vista: A visually, socially, and temporally-aware model for artistic recommendation[C]//Proceedings of the 10th ACM conference on recommender systems. 2016: 309-316.

[11] Tang J, Wang K. Personalized top-n sequential recommendation via convolutional sequence embedding [C]//Proceedings of the eleventh ACM international conference on web search and data mining. 2018: 565-573.

[12] Yuan F, He X, Jiang H, et al. Future data helps training: Modeling future contexts for session-based recommendation[C]//Proceedings of The Web Conference 2020. 2020: 303-313.

[13] Liu Q, Zeng Y, Mokhosi R, et al. STAMP: short-term attention/memory priority model for session-based recommendation[C]//Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining. 2018: 1831-1839.

[14] Ying H, Zhuang F, Zhang F, et al. Sequential recommender system based on hierarchical attention network[C]//IJCAI International Joint Conference on Artificial Intelligence. 2018.

[15] FENG S S,LI X T,ZENG Y F,et al. Personalized ranking metric embedding for next new POI recommendation. Proceedings of the 24th International Conference on Artificial Intelligence (IJCAI'15). AAAI Press,2015 : 2069-2075.