Research on session recommendation system based on graph neural network

Yan Wang [1, a], Wei Yu [2, b], and Lianqiang Niu [1, c]

[1] School of Software, Shenyang University of Technology ,Shenyang,China

[2] Neusoft，Shenyang,China

[a] 1466471473@qq.com, [b] yuwei@neusoft.comm, [c] niulq@sut.edu.cn

Abstract. before the graph neural network session recommendation system in order to ensure that the model can not smooth, and ignore the project long distance dependence, and most of the recommended method ignore the project sequence relationship, usually the session in the last project as the user's final interest, and lead to the session understanding is not comprehensive. In view of the above problems, this paper proposes a graph neural network recommendation method with sequence network. The main idea of this method is to regard the session as a graph structure, in which the nodes represent the items that the user interacts with, and the edges represent the interaction between the user and these items, such as click, purchase, etc. Then, the graph neural network can learn the embedding vector of each behavior in the graph. Comparing these behavioral vectors with the corresponding item vectors shows the user's interest in different items in the current session, and a self-attention mechanism is introduced into the graph neural network to distinguish long-distance dependencies in the learning sessions without oversmoothing. The sequence network can then be used to introduce bidirectional gating recurrent units and soft attention mechanisms when the vector representation of the item is fused into the vector representation of the session. In addition, combining contrast learning techniques to improve the expression of long tail items and indirectly improve the effect of the session recommendation model. The experimental results show that the recommendation algorithm of adding sequence network can reduce the complexity and improve the performance.

Keywords: recommendation system, session recommendation, deep learning, graph neural network.

1. Introduction

With the popularity of the Internet and mobile devices, there is a surge in information, making it difficult for people to find effective information in the vast amount of data, which is called "information overload"[1].People need to spend more time and effort on finding and organizing huge amounts of information. The traditional recommendation system mainly has four algorithms: content-based recommendation[2-4], Recommendation based on collaborative filtering[5-7], Recommendation based on matrix decomposition[8-9], And recommendations based on implicit semantics[10-11]. However, because the information of many users is unknown, the traditional recommendation system is not highly accurate. Session-based recommendations have emerged. The conversation structure can avoid the loss of local information to the greatest extent, pay more attention to the dynamic and local parts of the session data, and be more in line with the current development environment of online services. With the development of session recommendation system, the method based on deep neural network has become the mainstream trend, and the method based on graph neural network is the most prominent. This study analyzes the current mainstream session recommendation system, and proposes the better conversation recommendation system based on graph neural network.

2. Organization of the Text

2.1 recurrent neural network

Circulating neural networks have achieved great success in sequence modeling and are tried to be applied in session recommendations[12-14], And achieved better performance than the conventional methods. GRU is the recurrent neural network that is most used in session recommendations[15-17], The underlying recurrent neural network RNN was applied in earlier methods based on recurrent neural networks[18], While the newer recurrent-based neural methods use long, short-term memory networks[19-21]. The method based on recurrent neural network has a rough transformation relationship between projects. It is believed that there is only a sequence relationship between projects, and it is difficult to learn the long distance dependence between projects. It is believed that there is a strict order relationship between projects, which leads to the failure of parallel calculation during model training, which increases the operation time of the model.

2.2 attention mechanism

Attention mechanism has excellent parallel properties. Early attention mechanisms in image captioning[21],, machine translation[22], And other natural language processing tasks achieved successful. In general, the use of attention mechanism in session recommendation tasks consists of two steps: focus on weight calculation and vector aggregation. The attention mechanism is used not only for dependent learning between projects, but also for the extraction of user interests. Although the attention mechanism-based method can learn the long distance dependence between items in a session, the proximity dependence and the long distance dependence cannot be well distinguished, and because the attention mechanism-based method regards a session as a disordered set, this ignores the sequence relationship between items.

2.3 graph neural network

In the session recommendation, the graph neural network types can be divided into three types: gated graph neural network, graph convolutional neural network, and graph attention neural network. The research framework of GNN-based session recommendation method is shown in Figure 1, similar to most graph neural network methods, which can be divided into three modules according to the corresponding functions: (1) graph neural network for the vector representation of learning projects; (2) sequence network, that is, user interest extraction network, to extract the vector representation of user interest learning sessions; and (3) prediction layer for calculating the recommendation score of the items to be recommended.

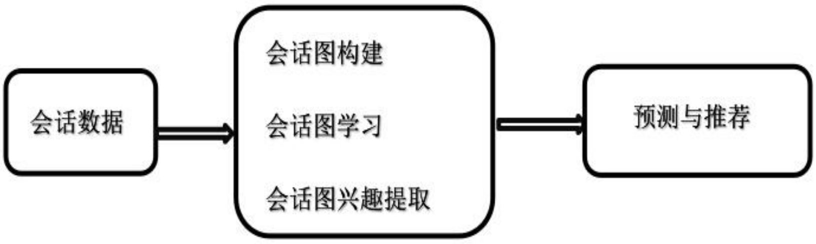

Figure 1 Research framework of the session recommendation method based on

In this paper, a sequence network is added to use the sequence information of the project in the session to aggregate the vector representation of the project. First, the self-attention mechanism is used for the item vector in the session, not only taking the last item of the session as a user representation. Moreover, without simply adding each line of elements of the attention weight matrix, sequence information is needed to measure how well each item reflects the user's interest. To this end, this paper introduces bidirectional gating cycle units to capture the sequence information of each item in the session:

$$h^{(i)}=\mathrm{GRU}(x_i^{(l)}, h^{(i-1)}), \quad (1)$$

GRU is a bidirectional gating cycle unit (GRU), $x_i^{(l)} \in \mathbb{R}^{d}$ is the ith vector representation of the first project in the session sequence in the graph neural network. $h^{i} \in \mathbb{R}^{2d}$ is the hidden state of GRU and contains the sequence information of the project in the session. The initial value of the GRU hidden state is set to the vector of all 0. The GRU's last hidden state does not represent the sequence of information for the entire session.Therefore, the sequence information of the entire session is obtained by averaging each hidden state:

$$h^{'} = \frac{1}{n}\sum_{i=1}^{n} h^{(i)}, \quad (2)$$

Next, the soft attention mechanism is used to dynamically measure the interest response of the project according to the sequence information of each item:

$$\gamma = q_2^{T}\sigma(W_1 H^{T} + W_2 h^{'} + b), \quad (3)$$

W_1, $W_2 \in \mathbb{R}^{d\times 2d}$ and W_1, $W_2 \in \mathbb{R}^{d\times 2d}$ is a learnable parameter,W_1, $W_2 \in \mathbb{R}^{d\times 2d}$ is a series of hidden states of the GRU cell, γ is the sequence reflection score for each item.These scores can then be used to measure how well all items affect each item:

$$\delta = B\gamma^{T}, \quad (4)$$

Measure the impact of each project on all projects based on their sequence information and add them up.Sequence information is lost if δ's each element is aggregated to form a session vector based on the degree to which the item reflects the user's interest, because the item in the session reflects the overall user's interest and its sequence information in the session is related.Therefore, it is also necessary to exercise weight control over δ's the elements of the impact score pair of each item:

$$\epsilon\delta\gamma^{T} \quad =\odot, \quad (5)$$

The practice uses the sequence information of the item in the session to dynamically measure how well each item reflects on user interest. With the degree to which each item reflects the user's interest, a session vector representation can be generated:

$$\mathrm{S}=[v_1, \ldots, v_n]^{T}\epsilon, \quad (6)$$

Where v is the vector representation of each item. he session vector S consists of all the items in it. Through the action of the graph neural network and the sequence perception network, the corresponding weight of each item vector is related to the transformation of the adjacent items and the long distance of the items, as well as the sequence information in the session.

3. Experimental design and analysis of the results

3.1 Experimental process flow

Filter out less than 5 items in the entire dataset, and then delete sessions with less than 2 session lengths. The session of the first 6 days of the data set was used as the training set and the last day session as the test set. The embedding vector dimension of the item was set to d=150 and batchsize chosen to 150, and all parameters of the model were initialized using a Gaussian distribution with a mean of 0.1 and a standard deviation of 0.2. The initial learning rate was 0.002, with a decay of 0.2 after every 3 epochs. To evaluate the effectiveness of the S-GNN sequence network, it will be compared with the three classical models.

3.2 Evaluation indicators

In the evaluation of the model index, with previous work[22-23]The most common P@k and MRR @ K are consistently adopted. The calculation methods of the two indicators are introduced below.

P@K (Precious): The score of P@K measures whether the target item is included in the first K items of the recommended item list, which is generally ranked by the recommendation score.

$$P@K = \frac{n_{hit}}{N} \tag{7}$$

n_{hit} Where N is the number of session sequences in the test set and the number of times the target item is located in the K item before the ranked list. Its evaluation is relatively loose, and it does not have a strong correlation with the recommended score of the target project.

MRR @ K (Mean Reciprocal Rank): MRR @ K takes into account the position of the target project in the recommended list. If the recommended score of the target project is not in the first K, the recommended score is 0. Otherwise, it is calculated as follows:

$$MRR@K = \frac{1}{N} \sum_{v_{target} \in S_{test}} \frac{1}{Rank(v_{target})} \tag{8}$$

v_{target} Where is the target items, Rank () is the ranked position of the target items in the recommended list. MRR @ K versus P@K is a standardized recommended hit score that considers the location of the target item.

3.3 Experimental results and analysis

The session length measures the sequence generation layer has a great impact, so two datasets with longer session length are selected to compare SR-GNN, NISER and GCE-GCN. The comparison results are shown as follows below:

- GNN-NP: The sequence network of S-GNN is replaced with the SR-GNN user interest extraction network, which is the most used user interest extraction strategy in the graph neural network method.
- GNN-PP: The sequence network of S-GNN is replaced by the network of user interest extraction network of NISER, which adds a forward position embedding vector to the user interest extraction network of SR-GNN.
- GNN-RP: The sequence network of S-GNN is replaced by the session vector generation network of GCE-GNN, which uses the reverse position encoding.

Table 1 Performance comparison of different user interest extraction networks

Datasets	Metrics	GNN-NP	GNN-PP	GNN-RP	S-GNN
Diginetica	P@20	51.26	53.55	53.65	55.05
	MRR@20	18.26	18.25	18.55	19.85

It can be seen that the performance effect of GNN-NP is the worst, which indicates that it is not enough to focus on the last item in the session. GNN-PP improves the model efficiency by using positive absolute position coding to prevent different sessions from generating the same session vector problem. But absolute position coding does not distinguish relative position problems between items. GNN-RP improved relatively little, probably because of the session length of the dataset. The sequence network has the best performance compared to the other three session vector layers, demonstrating its effectiveness and confirming that the sequence network can accurately measure the importance that each item in the session reflects the interest in the user.

4. Conclusion

Recommendation recommendation is one of the new areas to help retain users, when traffic is king. Unlike the recurrent neural network based approach that treats the items in a session as having a strict sequential relationship, or the attention mechanism based approach which treats the whole session as a set of unsequential relationships, the graph neural network approach is also more able to learn the correlation between items. Since only one data set was selected in this experiment, multiple datasets were selected for trial, and the graph neural network layer was not compared, the following will be compared to further optimize the model.

References

[1] Wang S, Cao L, Wang Y, etal.Asurveyon session-based recommender systems[J].ACM Computing Surveys, 2021, 54(7): 1-38.

[2] Lops P, De Gemmis M, Semeraro G.Content-based recommender systems: State of the art and trends[M].Recommender Systems Handbook, 2011: 73-105.

[3] Pazzani M J, Billsus D.Content-based recommendation systems[J].The Adaptive Web: Methods and Strategies of Web Personalization, 2007, 4321: 325-341.

[4] Lu H, Ma W, Zhang M, et al.Standing in your shoes: External assessments for personalized recommender systems[C].Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval.2021: 1523-1533.

[5] Schafer J B, Frankowski D, Herlocker J, et al.Collaborative filtering recommender systems[J].The Adaptive Web: Methods and Strategies of Web Personalization, 2007, 4321: 291-324.

[6] Ekstrand M D, Riedl J T, Konstan J A.Collaborative filtering recommender systems[J].Foundations and Trends in Human-Computer Interaction, 2011, 4(2): 81-173.

[7] Pi Q, Zhou G, Zhang Y, et al.Search-based user interest modeling with lifelong sequential behavior data for click-through rate prediction[C].Proceedings of the 29th ACM International Conference on Information & Knowledge Management.2020: 2685-2692.

[8] Zheng V, Cao B, Zheng Y, et al.Collaborative filtering meets mobile recommendation: A user-centered approach[C].Proceedings of the AAAI Conference on Artificial Intelligence, 2010, 24(1): 236-241.

[9] Chen S, Peng Y.Matrix factorization for recommendation with explicit and implicit feedback[J].Knowledge-Based Systems, 2018, 158: 109-117.

[10] Luo X, Zhou M C, Li S, et al.A nonnegative latent factor model for large-scale sparse matrices in recommender systems via alternating direction method[J].IEEE Transactions on Neural Networks and Learning Systems, 2015, 27(3): 579-592.

[11] Wu D, Luo X, Shang M, et al.A deep latent factor model for high-dimensional and sparse matrices in recommender systems[J].IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2019, 51(7): 4285-4296.

[12] HIDASI B,KARATZOGLOU A,BALTRUNASL,et al.Session-based Recommendations with Recurrent Neural Networks[C/OL]//4th International Conference on Learning Representations, Conference Track Proceedings.2016.http://arxiv.org/abs/1511.06939.

[13] LIJ,REN P,CHEN Z,et al.Neural Attentive Session-based Recommendation[C]//Proceedings ofthe 2017 ACM on Conference on Information and Knowledge Management.2017:1419-1428.

[14] WANG M,REN P,MEIL,et al.A Collaborative Session-based Recommendation Approach with Parallel Memory Modules[C]//Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval.2019:345-354.

[15] QUADRANA M,KARATZOGLOU A,HIDASI B,et al.Personalizing Session-based Recommendations with Hierarchical Recurrent Neural Networks[C]//Proceedings of the Eleventh ACM Conference on Recommender Systems.2017:130-137.

[16] RUOCCO M,SKREDE OS L,LANGSETH H.Inter-Session Modeling for Session-Based Recommendation[C]//Proceedings of the 2nd Workshop on Deep Learning for Recommender Systems.2017:24-31.

[17] 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.

[18] YU F,LIU Q,WU S,et al.A Dynamic Recurrent Model for Next Basket Recommendation[C] //Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval.2016:729-732.

[19] DOBROVOLNYM,SELAMATA,KREJCAR O.Session Based Recommendations Using Recurrent Neural Networks -Long Short-Term Memory[C]//Lecture Notes in Computer Science:Intelligent Information and Database Systems -13th Asian Conference,Proceedings:vol.12672.2021:53-65.

[20] GUO L,YIN H,WANG Q,et al.Streaming Session-based Recommendation[C]//Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery &Data Mining.2019: 1569-1577.

[21] LENZ D,SCHULZE C,GUCKERT M.Real-Time Session-Based Recommendations Using LSTM with Neural Embeddings[C]//Lecture Notes in Computer Science:Artificial Neural Networks and Machine Learning-27th International Conference on Artificial Neural Networks,Proceedings,Part II:vol.11140.Springer,2018:337-348.

[22] WUS,TANG Y,ZHU Y,et al.Session-Based Recommendation with Graph Neural Networks[C]// Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence.2019:346-353.

[23] WANG Z,WEI W,et Al.Global Context Enhanced Graph Neural Networks for Session-based Recommendation[C]//Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval.2020:169-178.