BERT-RS: A Neural Personalized Recommender System with BERT

Kezhi Lu, Qian Zhang, Guangquan Zhang and Jie Lu
Australian Artificial Intelligence Institute,
Faulty of Engineering and Information Technology,
University of Technology Sydney,
Sydney, NSW, 2007, Australia
E-mail: lukezhi@bjtu.edu.cn; {qian.zhang-1, guangquan.zhang, jie.lu}@uts.edu.au

Accurate user preferences and item representations are essential factors for personalized recommender systems. Explicit feedback behaviors, such as ratings and free-text comments, are rich in personalized preference knowledge and emotional evaluation information. It is a direct and effective way to obtain individualized preference and item latent representations from these sources. In this paper, we propose a novel neural model named BERT-RS for personalized recommender systems, which extracts knowledge from textual reviews and user-item interactions. First, we preliminary extract the semantic representation for users and items from the textual comments based on BERT. Next, these semantic embeddings are used for user and item latent representations through three different deep architectures. Finally, we carry out personalized recommendation tasks through the score prediction based on these representations. Compared with other algorithms, BERT-RS demonstrates outstanding experimental performance on the Amazon dataset.

Keywords: Collaborative Filtering; Recommender Systems; BERT; Personalized Recommendation.

1. INTRODUCTION

In the information age, the recommender system (RS) is crucial to solving the problem of information overload in Internet services. Providing personalized recommendations is the central goal for the RS. Meanwhile, the core in the personalized RS is accurately identifying the user's preferences and linking them with items or online services[1]. To this day, the collaborative filtering (CF) method has been used to obtain and predict users' similar preferences through the past interaction information between users and items (e.g., click and rating), which has been successfully and widely used in RS[2].

CF-based algorithms recommend relevant products according to users' similar preferences[3–5]. The CF-based model generally has two key points: (1) construct the similarity representation between users and items according to the past interaction records; (2) construct the model to train and recommend relevant items. For example, the neural collaborative filtering (NCF) method uses nonlinear neural networks with different structures based on matrix factorization (MF) to construct the representation of users and items[3,6]. However, the NCF model can not deeply analyze the potential preference information and emotional factors according to textual comments. As a result, in this paper, we propose a novel model utilizing reviews knowledge and user-item interactions to construct user/item representations for personalized recommendations.

Review texts contain rich knowledge of personal preferences and sentiments, which can facilitate the accurate representation of users and items. In NLP, various BERT-based models show their state-of-the-art performance in multiple tasks such as Named Entity Recognition (NER) and Question Answering (QA)[7]. Inspired by this, we fuse the BERT with neural architectures to extract semantic embeddings from reviews content and construct the personalized user/item representations based on them for RS.

In general, BERT-RS has three different parts. Firstly, we extract and fine-tune the preliminary semantic embeddings from reviews based on the BERT. Secondly, we fuse semantic information to construct user and item personalized representations through three different neural architectures, i.e., BERT, GMF, and MLP. Lastly, we carry out personalized recommendation tasks through the score prediction based on these representations. The main contributions of this paper are summarized as follows:

- We propose BERT-RS, a novel method for personalized recommendations utilizing textual reviews based on BERT, GMF, and MLP, which shows that auxiliary knowledge of user reviews helps improve performance.
- Our proposed method not only captures the characteristic and emotional knowledge of users and items at the semantic level but also extracts the interactive knowledge between them.
- The outstanding experimental performance in the Amazon dataset compared with NCF proves the reliability of our method.

2. RELATED WORK

This section will briefly introduce related research work about CF-based and BERT model for recommender systems.

CF-based methods obtain and predict users' similar preferences through utilizing the past interactive information between users and items (e.g., click and rating), which is successfully and widely used in recommendation systems. For example, the Tapestry system (Goldberg et al., 1992) is one of the first industrial systems to utilize CF-based methods for the recommendation task[2]. Later, various CF-based algorithms emerged and were applied to the industrial RS. In particular, the matrix factorization (MF) method is further improved by mapping user-item representations into a shared latent space[8,9]. With the development of neural networks, NCF forms user-item representations by integrating CF with multi-layer perceptrons (MLP), which shows state-of-the-art performance on different datasets for RS[3]. In the meantime, in various NLP tasks, the pre-trained language representations model effectively improves their performance[10]. For example, Devlin et al. (2018) propose a novel Bidirectional Encoder Representations from Transformers (BERT) as a new language representation model[7], which demonstrates its state-of-the-art performance for various downstream NLP tasks, such as QA and NER. Recently, Qiu et al. (2021)[11] pre-trains user and item representations based on BERT for cross-domain recommendations. Inspired by these works, in this paper, we fuse the BERT model to extract the latent semantic embeddings in reviews for user/item representations and propose a new model for downstream recommendations.

3. METHODS: BERT-RS

BERT-RS includes five main components: Figure 1A mainly introduces the preprocessing steps of comment content; Figure 1B-1D illustrates three different nonlinear neural networks to construct personalized user/item representations. Figure 1E shows the final prediction layer of BERT-RS for the recommendation.

3.1. *Preprocessing Steps as Input Layer*

Given a comment text r, we mark its head with a token, [CLS], and its tail with [SEP]. As shown in Figure 1A, each word in the sentence is represented by position embeddings, segment embeddings, and token embeddings. These three embeddings of each word in the sentence are added and put into the layer normalization (LN) as the final embeddings. Finally, each review forms an embedded representation of $\boldsymbol{R}^{ui}_{l\times d}$. l represents the length of the review content, d represents the dimensions for each word embedding, and u, i represent the user and item respectively.

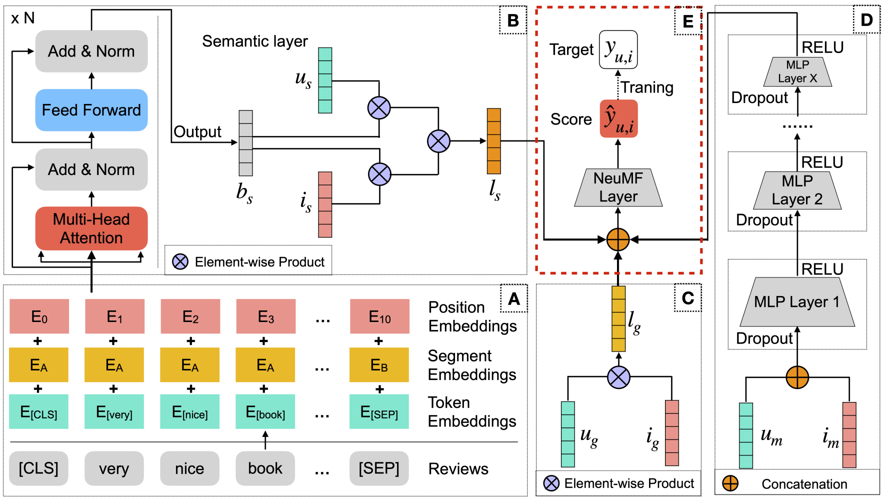

Fig. 1. Overview and framework of BERT-RS. (A) Preprocessing steps. (B) Semantic layer. (C) GMF layer. (D) MLP layer. (E) Prediction layer.

3.2. *Semantic Layer for User/Item Representations*

As illustrated in Figure 1B, this paper mainly utilizes the BERT model to extract the semantic information in the review. BERT model is mainly composed of n Transformer layers. $\boldsymbol{S}_k$ denotes the input embeddings in the k-th Transformer layer. It should be noted that $\boldsymbol{R}^{ui}_{l\times d}$ is the 0-th input embedding. As shown in Figure 1B, each Transformer layer has three key components, which are the Multi-Head Self-Attention layer, Add & Norm layer, and Feed-Forward layer.

3.2.1. *Multi-Head Self-Attention Layer*

In normal attention network, three different matrices are set as $\boldsymbol{Q}_{L_q\times d}$, $\boldsymbol{K}_{L_k\times d}$, and $\boldsymbol{V}_{L_v\times d}$ respectively ($L_k = L_v$). The formula is defined as:

$$Attention(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = Softmax(\boldsymbol{Q}\boldsymbol{K}^T/\sqrt{d})\boldsymbol{V} \tag{1}$$

Compared with the normal Attention layer, the Multi-head Self-Attention layer integrates multiple sub-layers to extract the Attention information, which is defined as:

$$Multi - Head - Attention(\boldsymbol{S}_k) = [head_1; ...; head_h]\boldsymbol{W}^* \tag{2}$$

$$head_j = Attention(\boldsymbol{S}_k\boldsymbol{W}^Q_j, \boldsymbol{S}_k\boldsymbol{W}^K_j, \boldsymbol{S}_k\boldsymbol{W}^V_j) \tag{3}$$

h represents the total number of Multi-heads. $\boldsymbol{W}_j^Q$, $\boldsymbol{W}_j^K$, and $\boldsymbol{W}_j^V$ represent $d \times d/h$ dimensional parameters. $\boldsymbol{W}^*$ represents the $d \times d$ dimensional parameters.

3.2.2. *Feed-Forward Layer*

We define this sub-layer as following:

$$FF(\boldsymbol{I}) = GELU(\boldsymbol{I}\boldsymbol{W}^{FF1} + \boldsymbol{b}^{FF1})\boldsymbol{W}^{FF2} + \boldsymbol{b}^{FF2} \tag{4}$$

where $\boldsymbol{I}$ represents the $L_i \times d$ dimensional input. $\boldsymbol{W}^{FF1}$, $\boldsymbol{b}^{FF1}$), $\boldsymbol{W}^{FF2}$, and $\boldsymbol{b}^{FF2}$ are parameters with $d \times d$, $4d \times d$, $4d$, and d dimension respectively. GELU represents the GELU activation function. The Add & Norm layer adds embeddings and performs LN, which is the same as the previous research work[12]. We define layer normalization as LN. The final output of the Transformer layer is defined as:

$$\boldsymbol{S}^{k+1} = LN(\boldsymbol{I}^k + FF(\boldsymbol{I}^k)) \tag{5}$$

$$\boldsymbol{I}^k = LN(\boldsymbol{S}^k + Multi - Head - Attention(\boldsymbol{S}^k)) \tag{6}$$

where $\boldsymbol{S}^{k+1}$ represents the final contextual semantic representation for reviews. Then this representation was inputted into a pooler layer, which consists of a linear layer and a Tanh activation function. As shown in Figure 1B, we define the output as $\boldsymbol{b}_s$. After that, we construct our semantic personalized user/item embedding ($\boldsymbol{u}_s$, $\boldsymbol{i}_s$) through element-wise product operation. Finally, we obtain the user and item semantic interaction embedding, $\boldsymbol{l}_s$. The detailed formula is defined as:

$$\boldsymbol{b}_s = Tanh(\boldsymbol{S}^{k+1}\boldsymbol{W}^{p1} + \boldsymbol{b}^{p1}) \tag{7}$$

$$\boldsymbol{l}_s = (\boldsymbol{u}_s \odot \boldsymbol{b}_s) \odot (\boldsymbol{i}_s \odot \boldsymbol{b}_s) \tag{8}$$

where $\odot$ represents the element-wise product.

3.3. *Neural User/Item Representations Based on GMF*

As illustrated in Figure 1C and inspired by NCF[3], we define the user latent embedding of GMF layer as $\boldsymbol{u}_g$ and item latent embedding as $\boldsymbol{i}_g$. The detailed function is defined as:

$$\boldsymbol{l}_g = \boldsymbol{u}_g \odot \boldsymbol{i}_g \tag{9}$$

3.4. *Neural User/Item Representations Based on MLP*

As can be shown in Figure 1D and inspired by NCF[3], we define the user latent embedding of MLP layer as $\boldsymbol{u}_m$ and item latent embedding as $\boldsymbol{i}_m$. This layer mainly consists of different MLP layers, which is defined as:

$$\boldsymbol{z}_1 = \phi_1(\boldsymbol{u}_m, \boldsymbol{i}_m) = [\boldsymbol{u}_m, \boldsymbol{i}_m]^T, \tag{10}$$

$$\phi_L(\boldsymbol{z}_{L-1}) = \alpha_L(\boldsymbol{W}_L^T \boldsymbol{z}_{L-1} + \boldsymbol{b}_L), \tag{11}$$

where $\boldsymbol{W}_L^T$, $\boldsymbol{b}_L$ and α_L represent the weight matrix, bias vector, and activation function respectively. We use the RELU function as the activation function for the MLP layer and define the output of MLP layer representation as $\boldsymbol{l}_m$.

3.5. *Prediction Layer*

As illustrated in Figure 1E, concerning the prediction layer, we concatenate the outputs of these three layers as the final embedding and utilize a linear layer to predict the target score, $\hat{y}_{ui}$, for the recommendation. Then, we use the sigmoid activation function as a probabilistic function to map $\hat{y}_{ui}$ in the range of [0, 1]. The final loss function, L, is defined as:

$$L = \frac{1}{n} \sum_{(u,i)\in Y} y_{ui} \times \boldsymbol{ln} \hat{y}_{ui} + (1 - y_{ui}) \times \boldsymbol{ln}(1 - \hat{y}_{ui}) \tag{12}$$

where Y denotes the interaction set of user, u, and item, i. n represents the total number of interactions.

4. EXPERIMENTS

4.1. *Datasets and Evaluation Metrics*

We take experiments based on the real-world and public Amazon dataset[13]. This dataset includes reviews and rating data that is suited for our task. Specifically, we randomly sample 10000 users with at least 20 comments per-user in the book category and pick each item with no less than ten reviews. In the training stage, we convert the interaction records between a user and an item to a value of 1. At the same time, we negative sample with the same number of user comments and set their interactive value to 0. Finally, we shuffle 70% of the interactive information as the training set and 30% as the test set. Table 1 shows the statistics of the final dataset.

Table 1. Statistics of the dataset.

Datasets	**Category**	**User**	**Item**	**Reviews**	**Sparsity**
Amazon	Books	3764	17647	83599	99.87%

We adopt two widely used top-K ranking metrics, i.e., Hit Ratio@K (HR) and Normalized Discounted Cumulative Gain@K (NDCG) as our final evaluation metrics[1]. In detail, we choose HR@K and NDCG@K with $K \in [1, 3, 5, 10]$ to show the evaluation results compared with baselines.

4.2. *Results*

In the semantic layer of our experiment, we use the pre-trained BERT-base-uncased model as our initial model. Then we fine-tune this model through our reviews the same as the original BERT step with MLM and NSP tasks[7]. The configuration parameters are the same as the BERT-base-uncased model using 12 Transformer layers and 768 hidden sizes. In the MLP layer, we use 3 MLP layers and 32 dimensions as predictive factors that are the same as the GMF layer. About the optimizer, we utilize Adam with a learning rate of 1×10^{-4}. Table 2 shows the final results following:

Table 2. Performance comparison on Amazon dataset.

Method	HR@10	NDCG@10	HR@5	NDCG@5	HR@3	NDCG@3	HR@1	NDCG@1
BERT-RS	**0.534**	**0.393**	**0.446**	**0.365**	**0.390**	**0.342**	**0.276**	**0.276**
NCF	0.460	0.315	0.355	0.281	0.298	0.258	0.202	0.202
MLP	0.440	0.286	0.336	0.252	0.269	0.225	0.164	0.164
GMF	0.438	0.289	0.333	0.255	0.274	0.231	0.171	0.171

Compared with the other methods, HR@K and NDCG@K evaluation results of BERT-RS were significantly improved by nearly 30%. In particular, in the HR@1 and NDCG@1 results, our method improved by 36.63%.

5. CONCLUSION

In this paper, we propose a novel neural personalized user/item representations method, named BERT-RS, making use of reviews content based on BERT, GMF, and MLP, which demonstrates that the auxiliary information of user reviews helps to improve recommendation. In the final experimental results, our method has significantly improved in both HR and NDCG compared with baselines, which proves the effectiveness of our method.

References

1. J. Lu, Q. Zhang and G. Zhang, *Recommender Systems: Advanced Developments* (World Scientific, 2020).
2. D. Goldberg, D. Nichols, B. M. Oki and D. Terry, Using collaborative filtering to weave an information tapestry, *Communications of the ACM* **35**, 61 (1992).
3. X. He, L. Liao, H. Zhang, L. Nie, X. Hu and T.-S. Chua, Neural collaborative filtering, in *Proceedings of the 26th International Conference on World Wide Web*, 2017.
4. Q. Zhang, W. Liao, G. Zhang, B. Yuan and J. Lu, A deep dual adversarial network for cross-domain recommendation, *IEEE Transactions on Knowledge and Data Engineering* (2021).
5. T. Wang and Y. Fu, Item-based collaborative filtering with bert, in *Proceedings of The 3rd Workshop on e-Commerce and NLP*, 2020.
6. J. Lu, J. Xuan, G. Zhang and X. Luo, Structural property-aware multilayer network embedding for latent factor analysis, *Pattern Recognition* **76**, 228 (2018).
7. J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, *ArXiv Preprint ArXiv:1810.04805* (2018).
8. A. Paterek, Improving regularized singular value decomposition for collaborative filtering, in *Proceedings of KDD Cup and Workshop*, 2007.
9. Y. Koren, Factorization meets the neighborhood: a multifaceted collaborative filtering model, in *Proceedings of 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2008.
10. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser and I. Polosukhin, Attention is all you need, *Advances in Neural Information Processing Systems* **30** (2017).
11. Z. Qiu, X. Wu, J. Gao and W. Fan, U-bert: Pre-training user representations for improved recommendation, in *Proc. of the AAAI Conference on Artificial Intelligence. Menlo Park, CA, AAAI*, 2021.
12. J. L. Ba, J. R. Kiros and G. E. Hinton, Layer normalization, *ArXiv Preprint ArXiv:1607.06450* (2016).
13. J. Ni, J. Li and J. McAuley, Justifying recommendations using distantly-labeled reviews and fine-grained aspects, in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.