

Item-based Collaborative Filtering with BERT

Yuyangzi Fu

eBay Inc.

yfuyu@ebay.com

Tian Wang

eBay Inc.

twang5@ebay.com

Abstract

In e-commerce, recommender systems have become an indispensable part of helping users explore the available inventory. In this work, we present a novel approach for item-based collaborative filtering, by leveraging BERT to understand items, and score relevancy between different items. Our proposed method could address problems that plague traditional recommender systems such as cold start, and "more of the same" recommended content. We conducted experiments on a large-scale real-world dataset with full cold-start scenario, and the proposed approach significantly outperforms the popular Bi-LSTM model.

1 Introduction

Recommender systems are an integral part of e-commerce platforms, helping users pick out items of interest from large inventories at scale. Traditional recommendation algorithms can be divided into two types: collaborative filtering-based (Schafer et al., 2007; Linden et al., 2003) and content-based (Lops et al., 2011; Pazzani and Billsus, 2007). However, these have their own limitations when applied directly to real-world e-commerce platforms. For example, traditional user-based collaborative filtering recommendation algorithms (see, e.g., Schafer et al., 2007) find the most similar users based on the seed user's rated items, and then recommend new items which other users rated highly. For item-based collaborative filtering (see, e.g., Linden et al., 2003), given a seed item, recommended items are chosen to have most similar user feedback. However, for highly active e-commerce platforms with large and constantly changing inventory, both approaches are severely impacted by data sparsity in the user-item interaction matrix.

Content-based recommendation algorithms calculate similarities in content between candidate

items and seed items that the user has provided feedback for (which may be implicit e.g. clicking, or explicit e.g. rating), and then select the most similar items to recommend. Although less impacted by data sparsity, due to their reliance on content rather than behavior, they can struggle to provide novel recommendations which may activate the user's latent interests, a highly desirable quality for recommender systems (Castells et al., 2011).

Due to the recent success of neural networks in multiple AI domains (LeCun et al., 2015) and their superior modeling capacity, a number of research efforts have explored new recommendation algorithms based on Deep Learning (see, e.g., Barkan and Koenigstein, 2016; He et al., 2017; Hidasi et al., 2015; Covington et al., 2016).

In this paper, we propose a novel approach for item-based collaborative filtering, by leveraging the BERT model (Devlin et al., 2018) to understand item titles and model relevance between different items. We adapt the masked language modelling and next sentence prediction tasks from the natural language context to the e-commerce context. The contributions of this work are summarized as follows:

- Instead of relying on unique item identifier to aggregate history information, we only use item's title as content, along with token embeddings to solve the cold start problem, which is the main shortcoming of traditional recommendation algorithms.
- By training model with user behavior data, our model learns user's latent interests more than item similarities, while traditional recommendation algorithms and some pair-wise deep learning algorithms only provide similar items which users may have bought.
- We conduct experiments on a large-scale e-

commerce dataset, demonstrating the effectiveness of our approach and producing recommendation results with higher quality.

2 Item-based Collaborative Filtering with BERT

As mentioned earlier, for a dynamic e-commerce platform, items enter and leave the market continuously, resulting in an extremely sparse user-item interaction matrix. In addition to the challenge of long-tail recommendations, this also requires the recommender system to be continuously retrained and redeployed in order to accommodate newly listed items. To address these issues, in our proposed approach, instead of representing each item with a unique identifier, we choose to represent each item with its title tokens, which are further mapped to a continuous vector representation. By doing so, essentially two items with the same title would be treated as the same, and can aggregate user behaviors accordingly. For a newly listed item in the cold-start setting, the model can utilize the similarity of the item title to ones observed before to find relevant recommended items.

The goal of item-based collaborative filtering is to score the relevance between two items, and for a seed item, the top scored items would be recommended as a result. Our model is based on BERT (Devlin et al., 2018). Rather than the traditional RNN / CNN structure, BERT adopts *transformer encoder* as a language model, and use attention mechanism to calculate the relationship between input and output. The training of BERT model can be divided into two parts: *Masked Language Model* and *Next Sentence Prediction*. We re-purpose these tasks for the e-commerce context into *Masked Language Model on Item Titles*, and *Next Purchase Prediction*. Since the distribution of item title tokens differs drastically from words in natural language which the original BERT model is trained on, retraining the masked language model allows better understanding of item information for our use-case. Next Purchase Prediction can directly be used as the relevance scoring function for our item collaborative filtering task.

2.1 Model

Our model is based on the architecture of BERT_{base} (Devlin et al., 2018). The encoder of BERT_{base} contains 12 Transformer layers, with 768 hidden units, and 12 self-attention heads.

2.1.1 Next Purchase Prediction

The goal of this task is to predict the next item a user is going to purchase given the seed item he/she has just bought. We start with a pre-trained BERT_{base} model, and fine-tune it for our next purchase prediction task. We feed seed item as sentence A, and target item as sentence B. Both item titles are concatenated and truncated to have at most 128 tokens, including one [CLS] and two [SEP] tokens. For a seed item, its positive items are generated by collecting items purchased within the same user session, and the negative ones are randomly sampled. Given the positive item set I_p , and the negative item set I_n , the cross-entropy loss for next purchase prediction may be computed as

$$L_{np} = - \sum_{i_j \in I_p} \log p(i_j) - \sum_{i_j \in I_n} \log(1 - p(i_j)). \quad (1)$$

2.1.2 Masked Language Model

As the distribution of item title tokens is different from the natural language corpus used to train BERT_{base}, we further fine-tune the model for the masked language model (MLM) task as well. In the masked language model task, we follow the training schema outlined in Devlin et al. (2018) wherein 15% of the tokens in the title are chosen to be replaced by [MASK], random token, or left unchanged, with a probability of 80%, 10% and 10% respectively. Given the set of chosen tokens M , the corresponding loss for masked language model is

$$L_{lm} = - \sum_{m_i \in M} \log p(m_i). \quad (2)$$

The whole model is optimized against the joint loss $L_{lm} + L_{np}$.

2.1.3 Bi-LSTM Model (baseline)

As the evaluation is conducted on the dataset having a complete cold-start setting, for the sake of comparison, we build a baseline model consisting of a title token embedding layer with 768 dimensions, a bidirectional LSTM layer with 64 hidden units, and a 2-layer MLP with 128 and 32 hidden units respectively. For every pair of items, the two titles are concatenated into a sequence. After going through the embedding layer, the bidirectional LSTM reads through the entire sequence and generates a representation at the last timestep. The MLP layer with logistic function produces the estimated

| Method | Prec @1 | Prec @10 | Recall @10 | NDCG @10 |
|------------------------------|---------|----------|------------|----------|
| Bi-LSTM | 0.064 | 0.029 | 0.295 | 0.163 |
| BERT _{base} w/o MLM | 0.263 | 0.057 | 0.572 | 0.408 |
| BERT _{base} | 0.555 | 0.079 | 0.791 | 0.669 |

Table 1: Result on ranking the item

probability score. The baseline model is trained using the same cross-entropy loss shown in Eq. 1.

2.2 Dataset

We train our models on an e-commerce website data. We collected 8,001,577 pairs of items, of which 33% are co-purchased (BIN event) within the same user session, while the rest are randomly sampled as negative samples. 99.9999% of entries of the item-item interaction matrix is empty. The sparsity of data forces the model to focus on *generalization* rather than *memorization*. The rationale would be further explained with the presence of the statistics of our dataset. Another 250,799 pairs of items are sampled in the same manner for use as a validation set, for conducting early stopping for training.

For testing, in order to mimic the cold-start scenario in the production system wherein traditional item-item collaborative filtering fails completely, we sampled 10,000 pairs of co-purchased items with the seed item not present in the training set. For each positive sample containing a seed item and a ground-truth co-purchased item, we paired the seed item with 999 random negative samples, and for testing, we use the trained model to rank the total of 1000 items given each seed item.

3 Results

The results of our evaluation are presented in Table. 1. We do not consider the traditional item-to-item collaborative filtering model (Linden et al., 2003) here since the evaluation is conducted assuming a complete cold-start setting, with all seed items unobserved in the training set, resulting in complete failure of such a model. Following the same reason, other approaches relying on unique item identifier (e.g. itemId) couldn't be considered either in our experiment. We believe its a practical experiment setting, as for a large-scale e-commerce platform, a massive amount of new items would be created every moment, and ignoring those items from the recommender system would be costly and inefficient.

We observe that the proposed BERT model greatly outperforms the LSTM-based model. When only fine-tuned for the Next Purchase Prediction task, our model exceeds the baseline by 310.9%, 96.6%, 93.9%, and 150.3% in precision@1, precision@10, recall@10, and NDCG@10 respectively. When fine tuning for the masked language model task is added, we see the metrics improved further by another 111.0%, 38.6%, 38.3%, and 64.0%.

From the experiment, the superiority of proposed BERT model for item-based collaborative filtering is clear. It is also clear that adapting the token distribution for the e-commerce context with masked language model within BERT is essential for achieving the best performance.

In order to visually examine the quality of recommendations, we present the recommended items for two different seed items in Table. 2. For the first seed item 'Marvel Spiderman T-shirt Small Black Tee Superhero Comic Book Character', most of the recommended items are T-shirts, paired with clothing accessories and tableware decoration, all having Marvel as the theme. For the second seed item 'Microsoft Surface Pro 4 12.3" Multi-Touch Tablet (Intel i5, 128GB) + Keyboard', the recommended items span a wide range of categories including tablets, digital memberships, electronic accessories, and computer hardware. From these two examples, we see that the proposed model appears to automatically find relevant selection criteria without manual specification, as well as make decisions between focusing on a specific category and catering to a wide range of inventory by learning from the data.

4 Summary

In this paper, we adapt the BERT model for the task of item-based recommendations. Instead of directly representing an item with a unique identifier, we use the item's title tokens as content, along with token embeddings, to address the cold start problem. We demonstrate the superiority of our model over a traditional neural network model in understanding item titles and learning relationships between items across vast inventory.

Acknowledgments

The authors would like to thank Sriganesh Madhvanath, Hua Yang, Xiaoyuan Wu, Alan Lu, Timothy Heath, and Kyunghyun Cho for their support and discussion, as well as anonymous reviewers for

their helpful comments.

References

- Oren Barkan and Noam Koenigstein. 2016. Item2vec: neural item embedding for collaborative filtering. In *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE.
- P Castells, S Vargas, and J Wang. 2011. Novelty and diversity metrics for recommender systems: choice, discovery and relevance. In *International Workshop on Diversity in Document Retrieval (DDR 2011) at the 33rd European Conference on Information Retrieval (ECIR 2011)*.
- Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM conference on recommender systems*, pages 191–198. ACM.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *Proceedings of the 26th international conference on world wide web*, pages 173–182. International World Wide Web Conferences Steering Committee.
- Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. 2015. Session-based recommendations with recurrent neural networks. *arXiv preprint arXiv:1511.06939*.
- Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature*, 521(7553):436–444.
- Greg Linden, Brent Smith, and Jeremy York. 2003. Amazon. com recommendations: Item-to-item collaborative filtering. *IEEE Internet computing*, (1):76–80.
- Pasquale Lops, Marco De Gemmis, and Giovanni Semeraro. 2011. Content-based recommender systems: State of the art and trends. In *Recommender systems handbook*, pages 73–105. Springer.
- Michael J Pazzani and Daniel Billsus. 2007. Content-based recommendation systems. In *The adaptive web*, pages 325–341. Springer.
- J Ben Schafer, Dan Frankowski, Jon Herlocker, and Shilad Sen. 2007. Collaborative filtering recommender systems. In *The adaptive web*, pages 291–324. Springer.

A Appendix

Seed: Marvel Spiderman T-shirt Small Black Tee Superhero Comic Book Character

| Title | Score |
|---|--------|
| Marvel Black Panther Character Men's Superhero Socks Black | 0.9999 |
| Marvel SPIDERMAN HOMECOMING Birthday Party Range Tableware Supplies Decorations | 0.9999 |
| Marvel Compression Leggings Superhero T-shirts Long Pants Spiderman Sportswear | 0.9999 |
| Venom T-Shirt Funny Baby Venom Spiderman Black T-Shirt for Men and Women | 0.9998 |
| SPIDERMAN HOMECOMING LOGO T-SHIRT | 0.9998 |
| Marvel Comics Books Avengers The Incredible Hulk Transforming Tee Shirt Black | 0.9998 |
| Mens Short Sleeve T Shirt Slim Fit Casual Tops Tee Batman Iron Man Spider-Man | 0.9998 |
| MEMORIES Stan Lee Marvel Superhero T-Shirt Artwork Spiderman Avengers Mens Top | 0.9998 |
| Marvel Spider-Man Superior Spider-Man Mens Black T-Shirt | 0.9998 |
| Marvel Spider-Man: Far From Home Red Glow Mens Graphic T Shirt | 0.9998 |

Seed: Microsoft Surface Pro 4 12.3" Multi-Touch Tablet (Intel i5, 128GB) + Keyboard

| Title | Score |
|--|--------|
| Microsoft Surface 2 1573 ARM Cortex-A15 1.7GHz 2GB 64GB SSD 10.8" Touch | 0.9998 |
| Microsoft Xbox Live 3 Month Gold Membership Instant Delivery, Retail \$24.99 | 0.9998 |
| Genuine OEM 1625 AC Power Charger Adapter For Microsoft Surface Pro 3 | 0.9997 |
| For Sale Microsoft Surface Pro LCD display Touch Screen Digitizer Glass Assembly | 0.9997 |
| 65W AC Adapter For Microsoft Surface Book/ Pro 4 Q4Q-00001 1706 15V 4A | 0.9997 |
| 4GB 8GB 16GB DDR3 1333 MHz PC3-10600 DIMM 240Pin Desktop Memory Ram | 0.9995 |
| Samsung Galaxy Tab A 10.1" Tablet 16GB Android OS - Black (SM-T580NZKAXAR) | 0.9994 |
| For Microsoft Surface Pro 4 1724 Touch Screen LCD Flex Cable Ribbon X937072-001 | 0.9993 |
| RSIM 12+ 2018 R-SIM Nano Unlock Card Fits for iPhone XS MAX/XR/X/8/ | 0.9993 |
| Intel Core i5-4570S Quad-Core Socket LGA1150 CPU Processor SR14J 2.90GHz | 0.9993 |

Table 2: Sample recommended items with seeds