Intelligent classification and personalized recommendation of E-commerce products based on machine learning

Kangming Xu[1a,*,†], Huiming Zhou[1b,†], Haotian Zheng[1c,†], Mingwei Zhu[2], Qi Xin[3]

[1a]Computer Science and Engineering, Santa Clara University, CA, USA
[1b]Computer Science, Northeastern University, CA, USA
[1c]Electrical & Computer Engineering, New York University, New York, NY, USA
[2]Computer Information System, Colorado state university, Fort Collins, CO, USA
[3]Management Information Systems, University of Pittsburgh, Pittsburgh, PA, USA

[*]Corresponding author: kangmingxu87 @gmail.com
[†]Kangming Xu, Huiming Zhou, and Haotian Zheng contributed equally to this work and should be considered as co-first authors.

Abstract. With the rapid evolution of the Internet and the exponential proliferation of information, users encounter information overload and the conundrum of choice. Personalized recommendation systems play a pivotal role in alleviating this burden by aiding users in filtering and selecting information tailored to their preferences and requirements. This paper undertakes a comparative analysis between the operational mechanisms of traditional e-commerce commodity classification systems and personalized recommendation systems. It delineates the significance and application of personalized recommendation systems across e-commerce, content information, and media domains. Furthermore, it delves into the challenges confronting personalized recommendation systems in e-commerce, including data privacy, algorithmic bias, scalability, and the cold start problem. Strategies to address these challenges are elucidated. Subsequently, the paper outlines a personalized recommendation system leveraging the BERT model and nearest neighbor algorithm, specifically tailored to address the exigencies of the eBay e-commerce platform. The efficacy of this recommendation system is substantiated through manual evaluation, and a practical application operational guide and structured output recommendation results are furnished to ensure the system's operability and scalability.

Keywords: personalized recommendation system, E-commerce, data privacy, BERT model.

1. Introduction

The personalized recommendation systems employed by companies like Amazon, Netflix, and various online businesses are vigorously promoted through articles, competitions, and other means. These recommendation systems have become exceedingly popular in the e-commerce sector, significantly boosting metrics such as UV, [1]PV, GMV click conversion, and order conversion by 400%-500% after their implementation. The distinct characteristics of commodity recommendation highlight significant differences between commodities and articles or news, with commodities possessing inherent transactional attributes. Inaccurate recommendations hinder user interaction, making it challenging for users to click, browse, and ultimately make purchases. [2] Typically, product recommendations rely on user preference models derived from browsing history, orders, shopping cart additions, searches, likes, favorites, comments, and other behaviors. These models are trained offline using logistic regression (LR) and calculated via map reduce. Therefore, the application of machine learning in e-commerce has broad prospects and provides huge development opportunities for e-commerce enterprises. First, machine learning can realize personalized recommendation systems by analyzing massive user data and product information, thereby improving users' shopping experience and purchase conversion rate. Secondly, machine learning can also help e-commerce enterprises optimize marketing strategies, achieve precision marketing and pricing strategies, and increase sales and profits.

2. Related work

2.1. Traditional e-commerce commodity classification

The main goal of the traditional e-commerce commodity classification system is to divide and classify the goods according to certain rules or standards according to static information such as the attributes, categories and labels of the goods, so that users can easily find and browse the goods in different categories. Although the traditional e-commerce commodity classification system can help users quickly locate the category of required goods to a certain extent, it also has some shortcomings. One of them is its relatively low efficiency, especially in the face of large-scale goods and user data, the system's response speed can become slower. [3-5]This is mainly because traditional classification systems usually rely on static commodity attribute and category information for classification, and this information may need to be constantly updated and maintained, resulting in bottlenecks in the system when processing data. In addition, the traditional e-commerce commodity classification system may also have inaccurate or imperfect classification problems. [6] Although the traditional e-commerce commodity classification system provides the convenience of commodity display and browsing to a certain extent, its shortcomings such as low efficiency and inaccurate classification still exist, which need to be further optimized and improved to enhance user experience and system performance.

2.2. Personalized recommendation system

The Personalised Recommendation System is an advanced business intelligence platform designed to provide users with personalised information services and decision support by analysing users' interests, purchasing behaviour and other relevant data. Personalised recommendation systems not only improve user satisfaction and company sales, but also help solve the problem of information overload and help users find those long-tail goods that may have been overlooked. At present, personalized recommendation system has become a key technology in the fields of e-commerce, social media and digital entertainment. Personalized recommendation system can usually be applied to e-commerce, content information, video, advertising and other recommendation scenarios.

In e-commerce, there is a key challenge of information overload. [7] Personalized recommendation systems can effectively improve user filtering efficiency and thus address this issue. Content information recommendation is an important application area of personalised recommendation system. Through the personalised recommendation system, it can solve the two major problems of low user activity and difficult retention of new users, and help users discover new content in time, reduce user reading fatigue, and increase user dwell time. According to the latest data, in 2024, the monthly active users of Internet integrated video have reached 842 million, while the monthly active users of popular products such as Douyin short video have reached 400 million, and the monthly active users of [8]Kuaishou short video have also reached 250 million. In this huge user group, personalized recommendation system plays a key role.

Therefore, the emergence of personalized recommendation system has overcome the static information limitations of traditional recommendation system, and has brought great progress for e-commerce, content information and media industries. In summary, the personalized recommendation system has brought more efficient and intelligent services to various industries, and has become an important driving force to promote the development of the industry.

2.3. E-commerce personalized recommendation system and challenges

Despite their benefits, e-commerce personalized recommendation systems face several challenges. One of the primary challenges is data privacy and security concerns. Another challenge is algorithmic bias and fairness. Personalized recommendation algorithms may inadvertently reinforce biases or discrimination based on factors such as race, gender, or socioeconomic status. Scalability is a significant challenge for e-commerce personalized recommendation systems. As e-commerce platforms grow and accumulate more users and products, the volume of data to process and the computational resources required for personalized recommendations increase exponentially. Additionally, the cold start problem poses a challenge for e-commerce personalized recommendation systems, especially for new users, new products, and niche categories. Without sufficient historical data, it can be challenging to provide accurate and relevant recommendations, leading to a suboptimal user experience.

3. Methodology

BERT has been popular in the nlp field since google announced its excellent performance in 11 NLP tasks at the end of October 2018. In this article, we will examine the role and advantages of BERT model in e-commerce personalized recommendation system, and understand how it works through the model used by eBay. The advantage of BERT model in personalized recommendation system lies in its strong semantic understanding ability.

3.1. BERT Model Introduction

BERT is an algorithmic model that has broken records for a number of natural language processing tasks. The training of BERT Model can be divided into two parts: Masked Language Model and Next Sentence Prediction. In this algorithm, we migrate these tasks to e-commerce recommendation tasks: Masked Language Model on Item Titles and Next Purchase Prediction.

Masked Language Model on Item Titles: [9]The distribution of tokens in the project title on the e-commerce platform is very different from the distribution of tokens in the natural language corpus used for BERT model pre-training. In order to better understand the semantic information in the e-commerce recommendation environment, we take this task as one of the goals to retrain the model. In the training process, we follow Devlin et al. 's scheme and MASK 15% tokens in the project title, then the loss function of this part is as follows:

$$\mathrm{L_{lm}} = \sum_{\mathrm{m_i} \in \mathrm{M}} \mathrm{logP}(\mathrm{m_i}) \tag{1}$$

Next Purchase Prediction: In the BERT model, the Next Sentence Prediction is used to predict whether A sentence A is the last sentence of another sentence [10]B. We turn this into A recommendation problem: given a seed item A, predict whether another item B is the next item the user will click to buy. Replace sentence A in the original model with the title of item A and sentence B in the original model with the title of item B, concatenating the titles as input to the model. For seed items, the items purchased in the same user session are taken as positive examples, and negative examples are randomly selected from inventory. Thus, given the set of positive cases and the set of negative cases, the loss function of this part is:

$$\mathrm{L_{np}} = -\sum_{\mathrm{i_j} \in \mathrm{I_n}} \mathrm{logp}(\mathrm{i_j}) - \sum_{\mathrm{i_j} \in \mathrm{I_n}} \mathrm{log}(1 - \mathrm{p}(\mathrm{i_j})) \tag{2}$$

The joint loss function of the two parts of the model is as follows:

$$\mathrm{L_{l_m}} + \mathrm{L_{np}} \tag{3}$$

3.2. Experimental design

In this notebook, we adhere to the CRISP-DM (Cross-Industry Standard Process for Data Mining) pipeline, aiming to develop a recommendation system tailored for eBay's customers. The objective stems from addressing the challenge posed by eBay's extensive product offerings, which could overwhelm users and potentially deter them from completing purchases. [11] The resources include images of [12]products, metadata for each product and customer, and transaction data. The transactions_train.csv file contains product-customer entries that are not unique, meaning there can be multiple transactions for the same product by the same customer. Success in this project will be measured by the Mean Average Precision (MAP@12) metric. The higher the MAP@12, the better the recommendation system performs. Missing customer_ids in transactions_train.csv will be excluded from scoring. Additionally, we will manually test if the model recommendations make sense.

3.3. Data sets and preprocessing

The impact of missing values is further investigated, particularly in the `detail_desc` column, by examining the distribution of missing descriptions across different categories (`index_name`).

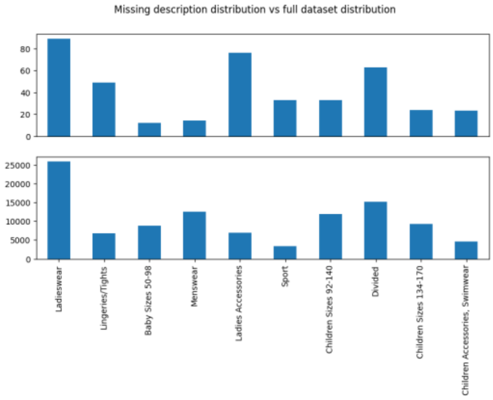

Figure 1. Histogram of distribution of the original data set

After comparing the histograms of descriptions with and without missing values, it's evident that the discrepancies in the counts of all index_names are minimal. The maximum number of missing entries per index_name is approximately 90, whereas even the smallest index_name (like "Sport") has around 2500 entries. Therefore, the decision has been made to utilize the entries without descriptions, despite constituting a small subset of the entire dataset. [13]This choice is grounded in the understanding that real-world scenarios may entail products without descriptions that could still be suitable for customers. Hence, relying on other categorical features is deemed sufficient.

In the preprocessing step, the lengths of sequences, such as product names and descriptions, were analyzed. The NaN values in the "detail_desc" column were filled with empty strings, and a function was devised to compute the length of text sequences. Multiple whitespaces were substituted with a single whitespace, and surrounding whitespace was removed before determining the sequence length. Subsequently, the distribution of description lengths was visualized using a histogram to gain insights into the data.

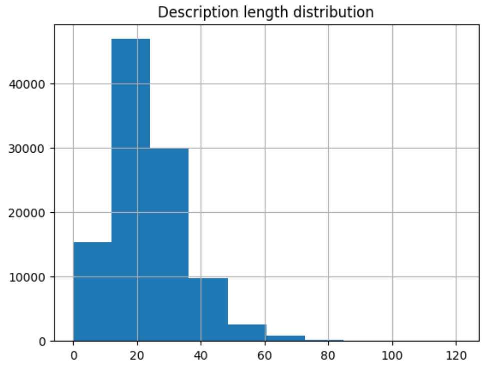

Figure 2. Data distribution model

3.4. Modeling

This section explains how to convert text into vector representations using the[14] BERT model, and uses the nearest neighbor algorithm in Sklearn to recommend products to customers. It first uses BERT to embed the text as a vector, then uses the nearest neighbor algorithm to find other products similar to the customer's purchase behavior based on their previous purchases, and finally provides the customer with a list of recommended products.

Table 1. Using all customer IDs.

	customer_id	prediction
0	00000dbacae5abe5e23885899a1fa44253a17956c6d1c3...	0706016001 0706016002 0372860001 0610776002 07...
1	0000423b00ade91418cceaf3b26c6af3dd342b51fd051e...	0706016001 0706016002 0372860001 0610776002 07...
2	000058a12d5b43e67d225668fa1f8d618c13dc232df0ca...	0706016001 0706016002 0372860001 0610776002 07...
3	00005ca1c9ed5f5146b52ac8639a40ca9d57aeff4d1bd2...	0706016001 0706016002 0372860001 0610776002 07...
4	00006413d8573cd20ed7128e53b7b13819fe5cfc2d801f...	0706016001 0706016002 0372860001 0610776002 07...

By comparing the purchase record of a specific customer with the system recommendation, it is found that the recommended product type is consistent with the customer's purchase preference. The system recommends a variety of different types of clothing, such as dresses and skirts, which coincide with a particular customer's purchase history. This indicates that the recommendation system can provide personalized recommendations according to customers' previous purchase behaviors, thus increasing customers' satisfaction and purchase intention.

Combining the results of the manual evaluation, use guidelines, and submission preparation, the following research conclusions can be expanded:

Through manual evaluation, the effectiveness of the recommendation system is confirmed, and it is found that the products recommended by the system are highly consistent with customers' purchase preferences, which provides strong support and verification for the recommendation algorithm. The user guide section describes in detail how to operate and use the system, providing clear guidance for practical applications, and users can easily use the system to provide personalized product recommendations for customers. [15]The submission preparation part shows the structured data output by the system, and provides the operation steps and sample code for submitting the recommendation results to the system, so that the research results can be successfully applied in practice and provide more intelligent and personalized recommendation services for e-commerce and other fields.

3.5. Experimental conclusion

By using BERT model for text vector representation and nearest neighbor algorithm for product recommendation, we designed a personalized recommendation system to solve the problem of eBay e-commerce platform. [16]Through this recommendation system, we can effectively improve the browsing experience of users and reduce the situation that users may give up buying because of search difficulties. At the same time, the recommendation system also helps to reduce unnecessary returns and associated transport emissions, thus having a positive impact on the environment. Through manual evaluation, we confirmed the effectiveness of the recommendation system and found that the types of products recommended were highly consistent with eBay users' purchase preferences. Combined with the usage guide and submission preparation, we provide clear operational guidance and structured output of recommendation results for [17]practical applications, thus ensuring the operability and scalability of the recommendation system. Therefore, our research results not only provide an effective solution, but also have feasibility and practicability in practice, which can provide more intelligent and personalized shopping recommendation service for eBay platform users.

4. Conclusion

In our proposed algorithm, each project is represented not by a conventional identifier but by a vector representation derived from the semantic understanding of its title tokens. By harnessing the semantic richness captured by the BERT model, our [18]algorithm transcends the limitations of traditional approaches, which often struggle to discern meaningful relationships between a large number of items. Through extensive experimentation conducted on real-world, large-scale datasets, we have empirically validated the efficacy of our algorithm. Our results demonstrate significant advantages in terms of understanding the nuanced semantic information inherent in items and effectively learning the intricate relationships between a multitude of projects.[19] To conclude, personalized recommendation system is expected to become the core technology in the field of e-commerce, bringing more business opportunities and competitive advantages to e-commerce enterprises.

References

[1] Zhang, Yuchen, et al. "Taxonomy discovery for personalized recommendation." Proceedings of the 7th ACM international conference on Web search and data mining. 2014.

[2] Tian, Yonghong, et al. "College library personalized recommendation system based on hybrid recommendation algorithm." procedia cirp 83 (2019): 490-494.

[3] Song, Xiaodan, et al. "Personalized recommendation driven by information flow." Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. 2006.

[4] Cheng, Qishuo, et al. "Optimizing Portfolio Management and Risk Assessment in Digital Assets Using Deep Learning for Predictive Analysis." arXiv preprint arXiv:2402.15994 (2024).

[5] Wu, Jiang, et al. "Data Pipeline Training: Integrating AutoML to Optimize the Data Flow of Machine Learning Models." arXiv preprint arXiv:2402.12916 (2024).

[6] Shepitsen, Andriy, et al. "Personalized recommendation in social tagging systems using hierarchical clustering." Proceedings of the 2008 ACM conference on Recommender systems. 2008.

[7] Zhou, Y., Tan, K., Shen, X., & He, Z. (2024). A Protein Structure Prediction Approach Leveraging Transformer and CNN Integration. arXiv preprint arXiv:2402.19095.

[8] Qian, Xueming, et al. "Personalized recommendation combining user interest and social circle." IEEE transactions on knowledge and data engineering 26.7 (2013): 1763-1777.

[9] Zhang, Denghui, et al. "E-BERT: A phrase and product knowledge enhanced language model for e-commerce." arXiv preprint arXiv:2009.02835 (2020).

[10] Wang, Yong, et al. "Construction and application of artificial intelligence crowdsourcing map based on multi-track GPS data." arXiv preprint arXiv:2402.15796 (2024).

[11] Ma, Haowei. "Automatic positioning system of medical service robot based on binocular vision." 2021 3rd International Symposium on Robotics & Intelligent Manufacturing Technology (ISRIMT). IEEE, 2021.

[12] Zheng, Jiajian, et al. "The Random Forest Model for Analyzing and Forecasting the US Stock Market in the Context of Smart Finance." arXiv preprint arXiv:2402.17194 (2024).

[13] Yang, Le, et al. "AI-Driven Anonymization: Protecting Personal Data Privacy While Leveraging Machine Learning." arXiv preprint arXiv:2402.17191 (2024).

[14] Zuo, Yi, et al. "Personalized recommendation based on evolutionary multi-objective optimization [research frontier]." IEEE Computational Intelligence Magazine 10.1 (2015): 52-62.

[15] Lei, Yaguo, Zhengjia He, and Yanyang Zi. "Application of an intelligent classification method to mechanical fault diagnosis." Expert Systems with Applications 36.6 (2009): 9941-9948.

[16] Liu, Na, et al. "A novel intelligent classification model for breast cancer diagnosis." Information Processing & Management 56.3 (2019): 609-623.

[17] Yacin Sikkandar, Mohamed, et al. "Deep learning based an automated skin lesion segmentation and intelligent classification model." Journal of ambient intelligence and humanized computing 12 (2021): 3245-3255.

[18] Chavez-Badiola, Alejandro, et al. "Embryo Ranking Intelligent Classification Algorithm (ERICA): artificial intelligence clinical assistant predicting embryo ploidy and implantation." Reproductive biomedicine online 41.4 (2020): 585-593.

[19] Haowei, M., Ebrahimi, S., Mansouri, S., Abdullaev, S. S., Alsaab, H. O., & Hassan, Z. F. (2023). CRISPR/Cas-based nanobiosensors: A reinforced approach for specific and sensitive recognition of mycotoxins. Food Bioscience, 56, 103110.