





ViEcomRec: A Dataset for Recommendation in Vietnamese E-Commerce

Quang-Linh Tran¹(✉) , Binh T. Nguyen², Gareth J. F. Jones¹ ,
and Cathal Gurrin¹

¹ ADAPT Centre, School of Computing, Dublin City University, Dublin 9, Ireland
linh.tran3@mail.dcu.ie, {gareth.jones,cathal.gurrin}@dcu.ie

² University of Science, Ho Chi Minh City, Vietnam
ngtbinh@hcmus.edu.vn

Abstract. Recent years have seen the increasing popularity of e-commerce platforms which have changed the shopping behaviour of customers. Valuable data from products, customers, and purchases on such e-commerce platforms enable the delivery of personalized shopping experiences, customer targeting, and product recommendations. We introduce a novel Vietnamese dataset specifically designed to examine the recommendation problem in e-commerce platforms, focusing on face cleanser products with 369,099 interactions between users and items. We report a comprehensive baseline experimental exploration into this dataset from content-based filtering to attribute-based filtering approaches. The experimental results demonstrate an enhancement in performance, with a 27.21% improvement in NDCG@10 achieved by incorporating a popularity score and content-based filtering, surpassing attribute-based filtering. To encourage further research and development in e-commerce recommendation systems using this Vietnamese dataset, we have made the dataset publicly available at https://github.com/linh222/face_cleanser_recommendation_dataset.

Keywords: Vietnamese datasets · e-commerce recommendation · content-based filtering

1 Introduction

With the development of the Internet and technological devices, e-commerce platforms have become hugely popular in recent years. The revenue generated through e-commerce continues to increase rapidly, showing significant growth during the COVID-19 pandemic, which imposed limitations on social interactions. Recommendations play a crucial role in this development to enhance customers' shopping experience and increase revenue from selling more products. While there is extensive research exploring methods for achieving reliable and effective recommendations, there are local features associated with the individual languages and markets of specific territories.

While there are a number of datasets available recording purchases and user behavior on e-commerce platforms such as Amazon¹ and other international e-commerce platforms [2], the availability of such datasets specific to Vietnamese remains limited for public use. The Vietnamese language poses challenges due to its complicated grammar structure and diverse word forms, making it difficult to analyze and process. Furthermore, the available resources for Vietnamese are limited and primarily focused on sentiment analysis [13, 15] and question answering [3]. Consequently, this paper introduces a novel dataset encompassing products, customers, and purchases from a Vietnamese e-commerce platform.

As well as describing this new dataset, we also report initial studies using this dataset. Content-based filtering [8] and attribute-based filtering is applied to run some initial experiments. In addition, the popularity of a product to customers is generally a significant factor influencing customers' purchase decisions. To assess its impact, an experiment is conducted to compare the performance by incorporating a popularity score.

2 Related Work

The recommendation problem has received significant attention from researchers due to its wide application in various domains, including food [14], and particularly e-commerce [10]. In e-commerce, the problem of recommendation has developed over the past few decades, starting with Ben Schafer's analysis [12] of six e-commerce platforms using recommender systems and the creation of a taxonomy of recommender systems in e-commerce.

Over time, more research has been conducted on recommendation systems in e-commerce, resulting in several benchmark datasets, especially in English-based e-commerce. The Amazon product reviews dataset [6] is one such benchmark dataset, consisting of customer reviews and ratings on Amazon from 1996 to 2014. Ahmed et al. [10] employed a context and attribute-aware cross-attention model to address next-item recommendations on four Amazon sub-datasets and achieved superior performance compared to previous systems. Other recommendation systems built upon this dataset, such as SSE-PT [17] proposed by Wu et al., utilized personalized transformers.

While there are numerous datasets recording activities for e-commerce platforms in English, the availability of datasets specific to Vietnamese e-commerce remains limited. Truong et al. [16] developed a recommendation database that incorporated customer preferences, purchase history, and 2,000 Vietnamese comments for employing opinion mining in recommendations. Nguyen et al. [9] examined the impact of online product recommendation systems on customer behavior on Vietnamese e-commerce websites. However, these studies did not introduce a suitable dataset for the recommendation problem in Vietnamese e-commerce.

Content-based filtering [8] is a classic recommendation algorithm. Numerous studies have employed content-based filtering in diverse domains.

¹ <https://www.amazon.com/>.

The e-commerce platforms, with their extensive product content, also present a promising application area for content-based filtering. Ruining et al. [2] highlighted the importance of product attributes, demonstrating that an attribute-aware recommendation system outperforms previous approaches.

3 Dataset

3.1 Dataset Crawling

In this study, data on face cleanser products is crawled from Shopee², a large e-commerce platform in Vietnam. Face cleansers attract significant attention from both men and women, resulting in many purchases on Shopee. We use Beautiful Soup³ and Selenium⁴ written in Python to crawl the face cleanser product information first and get the total number of 2244 items. From the information of items, we continue crawling the reviews indicating the interaction between users and items. It is worth noting that Shopee only allows customers to review products only after making a purchase. We recorded 369,099 reviews from 304,708 users collected with several attributes, including reviews, ratings, and date-time information.

3.2 Attribute Extraction

The descriptions of items are long paragraphs describing the content of products and other information. We perform an attribute extraction stage to get the useful attributes from the products. We extract 9 attributes from the description: item name, ingredient, product_feature, skin_type, capacity, design, brand, expiry, and origin. These attributes cover all of the aspects that users may typically want to know when purchasing an item. InstructGPT [7] released by OpenAI is a powerful tool that can automate a wide range of tasks and is used to extract the attribute from the description in this study. We provide some examples of extracted attributes from the descriptions first and give them to InstructGPT, and then ask it to perform the attribute extraction on all 2244 items. An annotator will double-check the extracted attributes from InstructGPT to ensure extraction accuracy and correct any wrong extraction. All the extracted and preprocessed data is published on the same repository.

4 Methodology

4.1 Problem Definition

A **next-item recommendation problem** comprises of a set of users $\mathcal{U} := \{1, 2, \dots, U\}$, a set of items $\mathcal{I} := \{1, 2, \dots, I\}$, and a sequence of users past interactions $\mathcal{D} := ((u_1, i_1), (u_2, i_2), \dots, (u_{N-1}, i_{N-1}), (u_N, i_N) \in (\mathcal{U} \times \mathcal{I}))$ of pairs of a

² <https://shopee.vn/>.

³ <https://www.crummy.com/software/BeautifulSoup/bs4/doc/>.

⁴ <https://www.selenium.dev/>.

user and items. Given the purchased items i_1 to i_{N-1} , the objective is to predict the next item i_N that user u_N is likely to purchase. The input and output of the recommendation can be formulated as follows:

- **Input:** A set of users \mathcal{U} , a set of items \mathcal{I} , and the set of past interactions \mathcal{D} .
- **Output:** A ranked list of items \mathcal{L} for a user sorted by the probability that the user will purchase.

Content-based filtering is combined with a popularity score, it is referred to as the ‘content-based filtering with popularity score’ problem. In this case, we have an embedded-description matrix $\mathcal{C} \in \mathbb{R}^{\mathcal{I} \times j}$ containing item description vectors, where j represents the dimension of the embedding matrix. On the other hand, the ‘attribute-based filtering’ problem involves an attribute matrix $\mathcal{A} \in \mathbb{R}^{\mathcal{I} \times j}$ containing item attribute vectors, with j representing the number of attributes associated with each item.

4.2 Content-Based Filtering with Popularity Score

Content-based filtering [8] (CB) is a conventional recommendation algorithm that leverages information obtained from previously purchased items to provide recommendations to customers. In this research, item descriptions are collected and preprocessed prior to performing content-based filtering. To overcome the challenges associated with processing Vietnamese reviews, this study adopts several preprocessing techniques proposed in [15], as their effectiveness has been demonstrated.

In this study, we employ four pre-trained models to extract embeddings from processed item descriptions. These models include TF-IDF, BLIP [4], PhoBERT [5], and OpenAI Ada2⁵. TF-IDF converts item descriptions into a matrix by assessing a term’s importance based on its frequency in a specific description (TF) and rarity across the entire corpus (IDF). PhoBERT is a cutting-edge Vietnamese language model known for its strong performance in various natural language processing (NLP) tasks, including sentiment analysis [15]. BLIP is a multi-modal model capable of understanding both visual and textual information. Since descriptions contain valuable product information, utilizing BLIP can get meaningful embeddings for measuring item similarity. Ada2 is an embedding model developed by OpenAI, known for its advanced NLP applications such as ChatGPT, making it a reliable source for generating rich semantic embeddings. Once the descriptions are embedded, a cosine similarity calculation is performed between the embeddings of candidate items and purchased items to generate a ranked list of candidates.

Item popularity can significantly influence customer decisions, which the study [1] addresses biases from the popularity of products. We investigate the impact of item popularity on the content-based filtering recommendation system, by applying the popular score (number of sold products) to calculate the final

⁵ <https://platform.openai.com/docs/guides/embeddings>.

relevance score using the formula 1 with adjustable experimental parameters α and β .

$$\text{Relevance_score} = \alpha * \text{Cosine_Similarity} + \beta * \text{Popularity_Score} \quad (1)$$

4.3 Attribute-Based Filtering

Attribute-based filtering suggests items by extracting specific attributes from purchased items to identify similar items that share similar attributes. We utilize the extracted attribute in Sect. 3.2 to perform the attribute-based filtering.

Attribute-based filtering is performed using Elasticsearch⁶, an open-source search engine. Two approaches of attribute-based filtering are employed: text-based and embedding-based. In the text-based approach, the textual attributes of all items are indexed within Elasticsearch. Subsequently, the attributes of purchased items are fed to Elasticsearch to calculate the BM25 score [11] to perform searches on each attribute. The scores obtained from these searches are then combined through a weighted average, generating a list of similar items and their relevance scores.

Embedding-based attribute filtering is similar to content-based filtering described in Sect. 4.2. The textual attributes undergo an embedding process using the OpenAI Ada2⁷ language model, chosen for its superior performance in content-based filtering. To perform recommendations based on purchased items, cosine similarity calculations are carried out between the attribute embeddings of the purchased items and those of all candidate items. The cosine similarities for each attribute are then combined using a weighted average approach, resulting in a ranked list of candidate items.

5 Experiment

5.1 Experimental Settings

We use the leave-one-out protocol for training, validation, and testing the recommendation systems, which has been widely used in previous research [2, 10, 17]. Each customer’s two most recent interactions are withheld for validation and testing purposes, while the remaining previous interactions are utilized for training. Table 1 presents statistics on the three sets: training, validation, and testing.

In this study, the values of α and β in formula 1 are set to 0.7 and 0.3, respectively. These parameters are selected based on the weight-turning experiment. In attribute-based filtering, the weights assigned to different attributes are determined as follows: 0.7 for the item name, 0.5 for the ingredient and product_feature, and 0.5 for the other attributes. These parameter settings are chosen through a grid-search evaluation process.

⁶ <https://www.elastic.co>.

⁷ <https://platform.openai.com/docs/guides/embeddings>.

Table 1. Dataset Statistics

Dataset	Users	Items	Interactions
Train	304708	2244	358591
Validation	3592	900	5254
Test	3592	862	5254

Table 2. Content-based filtering versus Content-based filtering with the Popularity score

Model	R@10	MRR@10	NDCG@10
CB-TFIDF	0.1196	0.0586	0.2152
CB-Phobert	0.063	0.0377	0.1278
CB-BLIP	0.0786	0.0397	0.1426
CB-Ada2	0.1209	0.0535	0.2041
CB-TF-IDF+Popularity	0.1411	0.0661	0.2391
CB-Phobert+Popularity	0.0969	0.0415	0.1501
CB-BLIP+Popularity	0.1004	0.0451	0.1673
CB-Ada2+Popularity	0.1644	0.0742	0.2721

To assess the performance of the recommendation system on the new dataset, three metrics are utilized: Recall top K, Normalized Discounted Cumulative Gain (NDCG), and Mean Reciprocal Rank (MRR). The ranked list is truncated at a threshold value of 10, as this is a typical length for rank lists in various recommendation system studies [10].

5.2 Experimental Results

The results of four embedding models (TF-IDF, PhoBERT, BLIP, and Ada2) in the context of content-based filtering on the test set are presented in Table 2. Notably, PhoBERT and BLIP embeddings yield inferior results compared to TF-IDF and Ada2. Despite being trained on a large amount of data, PhoBERT and BLIP may struggle due to the generalization of embeddings and their lack of domain-specific knowledge. By contrast, TF-IDF and Ada2 perform well across the test set, achieving NDCG@10 scores of 21.52% and 20.41% for content-based filtering and 23.91% and 27.21% for content-based filtering with the inclusion of popularity scores, respectively. This demonstrates a significant improvement over PhoBERT and BLIP. Furthermore, adding the popularity score to content-based filtering noticeably enhances performance, leading to an increase of up to 7% in NDCG@10. From the experimental results, it becomes apparent that the dataset poses challenges in accurately recommending the next item, as evidenced by the highest R@10 score of only 16.44% and an NDCG@10 of 27.21%. We can conclude that content-based filtering performs moderately on the dataset for

Table 3. Text-based (TB) vs Embedding-based (EB) Attribute filtering with the Popularity score

Model	Recall@10	MRR@10	NGCD@10
TB + All attributes	0.1305	0.0638	0.2282
TB + All attributes except name	0.1029	0.0549	0.1934
TB + All attributes except product_feature	0.1310	0.0607	0.2245
TB + All attributes except ingredient	0.1376	0.0661	0.2372
TB + All attributes except design & expiry	0.1305	0.0622	0.2249
TB + Item name, ingredient and product_feature	0.1181	0.0574	0.2062
EB + All attributes	0.1368	0.0759	0.2595
EB + All attributes except name	0.1128	0.0534	0.1927
EB + All attributes except product_feature	0.1371	0.0583	0.2209
EB + All attributes except ingredient	0.1432	0.0612	0.2291
EB + All attributes except design & expiry	0.1449	0.0620	0.2297
EB + Item name, ingredient, and product_feature	0.1093	0.0521	0.1811

Table 4. Experimental results of different recommendation systems on the dataset

Model	Recall@10	MRR@10	NDCG@10
Random	0.0045	0.0014	0.0061
Top Popular	0.0812	0.0224	0.089
CB-Ada2	0.1209	0.0535	0.2041
CB-Ada2+ Popularity	0.1644	0.0742	0.2721
Text-based Attribute Filtering	0.1376	0.0661	0.2372
Embedding-based Attribute Filtering	0.1368	0.0759	0.2595

the next-item recommendation, with results varying depending on the chosen embedding model. The results demonstrate that incorporating the popularity score significantly improves the performance of all content-based models.

To compare the performance of text-based and embedding-based attribute filtering, we conducted an experiment and performed an ablation study on the attributes. The results are presented in Table 3. Embedding-based attribute filtering outperforms text-based filtering in all metrics. This can be because embeddings with cosine similarity carry more meaningful information than text-based BM25 similarity. However, the difference between the two approaches is relatively small, with an NDCG@10 improvement of only around 2%. When we selectively remove certain attributes to assess their importance in the overall performance, it becomes evident that the item name is the most important attribute. Removing the item name attribute in text-based and embedding-based attribute filtering results in a decrease in NDCG@10 of 3.48% and 6.68%, respectively.

A comprehensive experiment was conducted on attribute-based filtering using the dataset. The results of all recommendation models are presented in Table 4. The content-based filtering with a popularity score achieves the highest performance with 16.44% Recall@10, 7.42% MRR@10 and 27.21% NDCG@10. The embedding-based attribute filtering achieved the best performance among the attribute-based filtering approaches, with a Recall@10 of 13.68%, MRR@10 of 7.59%, and NDCG@10 of 25.95%. Although the Recall@10 and NDCG@10 scores of attribute-based filtering are not as high as those of content-based filtering, the MRR@10 score of attribute-based filtering is slightly better. Attribute-based filtering is not as effective as content-based filtering. This suggests that the entire description contains more information and has a stronger influence on attracting customers to purchase than extracting specific attributes.

6 Conclusions and Future Work

In this paper, we introduced a novel dataset designed to recommend face cleansers on a Vietnamese e-commerce platform. The dataset comprises 369,099 reviews from 304,708 customers, covering 2,244 unique products. An attribute extraction phase is conducted to extract valuable information from the product descriptions, which enables item recommendation based on attributes.

Baseline experimental results using this dataset indicate that content-based filtering, when combined with the popularity score, achieves the highest performance with an NDCG@10 of 27.21%. Additionally, attribute-based filtering is applied to the new dataset, and a comparative analysis is conducted between text-based and embedding-based attribute filtering approaches.

In the future, we want to continue improving the performance of recommendation systems on the dataset by utilizing additional data and advanced recommendation algorithms.

Acknowledgements. This research was conducted with the financial support of Science Foundation Ireland at ADAPT, the SFI Research Centre for AI-Driven Digital Content Technology at Dublin City University [13/RC/2106_P2]. For the purpose of Open Access, the author has applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission.

References

1. Abdollahpouri, H., Burke, R., Mobasher, B.: Managing popularity bias in recommender systems with personalized re-ranking (2019)
2. He, R., McAuley, J.: VBPR: visual Bayesian personalized ranking from implicit feedback. In: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, AAAI 2016, pp. 144–150. AAAI Press (2016)
3. Le, K., Nguyen, H., Le Thanh, T., Nguyen, M.: VIMQA: a Vietnamese dataset for advanced reasoning and explainable multi-hop question answering. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference, pp. 6521–6529, Marseille, France (2022). European Language Resources Association

4. Li, J., Li, D., Xiong, C., Hoi, S.: Bootstrapping language-image pre-training for unified vision-language understanding and generation, Blip (2022)
5. Nguyen, D.Q., Nguyen, A.T.: PhoBERT: pre-trained language models for Vietnamese. In: Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 1037–1042 (2020). Association for Computational Linguistics
6. Ni, J., Li, J., McAuley, J.: 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), pp. 188–197, Hong Kong, China (2019). Association for Computational Linguistics
7. Ouyang, L., et al.: Training language models to follow instructions with human feedback (2022)
8. Pazzani, M.J., Billsus, D.: Content-based recommendation systems. In: Brusilovsky, P., Kobsa, A., Nejdl, W. (eds.) *The Adaptive Web*. LNCS, vol. 4321, pp. 325–341. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-72079-9_10
9. Nguyen, P., Tho L.D.: The effect of online product recommendation system on consumer behavior: Vietnamese e-commerce websites **10**, 1–24 (2021)
10. Rashed, A., Elsayed, S., Schmidt-Thieme, L.: Context and attribute-aware sequential recommendation via cross-attention. In: Proceedings of the 16th ACM Conference on Recommender Systems, RecSys 2022, pp. 71–80, New York, NY, USA (2022). Association for Computing Machinery
11. Robertson, S., Zaragoza, H.: The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retrieval*. **3**, 333–389 (2009)
12. Schafer, J.B., Konstan, J., Riedl, J.: Recommender systems in e-commerce. In: Proceedings of the 1st ACM Conference on Electronic Commerce, EC 2009, pp. 158–166, New York, NY, USA (1999). Association for Computing Machinery
13. Tran, L.Q., Van Duong, B., Nguyen, B.T.: Sentiment classification for beauty-fashion reviews. In: 2022 14th International Conference on Knowledge and Systems Engineering (KSE), pp. 1–6 (2022)
14. Tran, Q.L., Lam, G.H., Le, Q.N., Tran, T.H., Do, T.H.: A comparison of several approaches for image recognition used in food recommendation system. In: 2021 IEEE International Conference on Communication, Networks and Satellite (COMNETSAT), pp. 284–289 (2021)
15. Tran, Q.L., Le, P.T. D., Do, T.H.: Aspect-based sentiment analysis for Vietnamese reviews about beauty product on E-commerce websites. In: Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation, pp. 767–776, Manila, Philippines (2022). De La Salle University
16. Truong, Q.-D., Thi Bui, T.D., Nguyen, H.T.: Product recommendation system using opinion mining on Vietnamese reviews. In: Phuong, N.H., Kreinovich, V. (eds.) *Soft Computing: Biomedical and Related Applications*. SCI, vol. 981, pp. 313–325. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-76620-7_27
17. Wu, L., Li, S., Hsieh, C.J., Sharpnack, J.: SSE-PT: sequential recommendation via personalized transformer. In: Proceedings of the 14th ACM Conference on Recommender Systems, RecSys 2020, pp. 328–337, New York, NY, USA (2020). Association for Computing Machinery