A Comprehensive Recommendation System for Online Shopping Based on the Google Dataset

Dongming Chen

Ocean University of China, 266100, Shandong, China
Chendongming@stu.ouc.edu.cn

Abstract. With the upgrading of online shopping software, intelligent recommendation systems have emerged in various software and become the main recommendation engine of contemporary times. This article outlines the construction of an all-encompassing recommendation system, drawing inspiration from prior collaborative filtering concepts. The system undergoes training and testing using the Google Shopping dataset. Scoring and evaluation texts are treated as distinct components, later combined to create a versatile recommendation system. Post-testing, the system is capable of furnishing recommendations based on all available data without user input. Alternatively, it can offer personalized recommendations derived from users' purchase histories, providing a selection of the top 15 recommendations to align with contemporary shopping software prerequisites. The purchase index with a test result of 63.0% for the entire dataset indicates that it fully meets user purchasing needs, but the average purchase rate is only 7.99%, and there is still room for improvement in this case.

Keywords: Recommendation system, Natural language processing, Singular Value Decomposition

1 Introduction

With the continuous development of society and the increasing liveliness of the internet, shopping recommendation systems have become one of the most important functions of online shopping apps. In fact, there are millions of products on shopping websites nowadays, so it is difficult for people to find what they truly want and the price is suitable, and it is difficult to determine whether these things are in good condition. The emergence of recommendation systems has solved the problems of both users and merchants. For assisting merchants in targeted recommendations to users in need, improving product transaction rates, and reducing advertising costs, for ordinary users [1], they can quickly screen out the products they need and prioritize the products that are most advantageous to them. So the key point of a recommendation system is how to recommend suitable products and ensure that users purchase them [1].

The realm of recommendation systems has witnessed a remarkable evolution over the past few decades, driven by advances in Artificial Intelligence (AI) techniques. These systems have become ubiquitous in the daily lives, influencing the

B. H. Ahmad (ed.), *Proceedings of the 2023 International Conference on Data Science, Advanced Algorithm and Intelligent Computing (DAI 2023)*, Advances in Intelligent Systems Research 180,
https://doi.org/10.2991/978-94-6463-370-2_33

choices in online shopping, content consumption, and more. In this paper, a comprehensive exploration of the development process of AI in recommendation systems was provided, delving into the multifaceted methods and techniques that have shaped this field.Early recommendation systems relied on rudimentary methods such as collaborative filtering [2], content-based filtering, and popularity-based recommendations. These methods, though simple, laid the foundation for more sophisticated approaches that would follow.

In the past, many studies have developed recommendation systems based on different items, including Netflix movie recommendation systems [3], Amazon product recommendations [4], Spotify music recommendations, YouTube video recommendations, and e-commerce website recommendations [5]. There are many technologies that can be used as references and applications. The Netflix movie recommendation system uses collaborative filtering algorithms such as matrix decomposition, resulting in a recommendation list. This project also presents some of the results in this way. Spotify utilizes collaborative filtering [2] and Natural Language Processing (NLP) technology to analyze users' music preferences, playback history, and song characteristics, and recommend music to users. They also use deep learning models to improve music recommendations. Many e-commerce websites use various recommendation algorithms, including collaborative filtering, content-based filtering, temporal models, etc., to recommend products to users. These algorithms generate personalized recommendation lists based on users' browsing, searching, and purchasing behaviors. But the factors they consider are limited, and there is not an integrated system to calculate all the influencing factors.

In terms of algorithms, collaborative filtering algorithms and content-based recommendation algorithms are currently mainly used, as well as neural networks and reinforcement learning for processing. However, they only use partial information, while hybrid recommendation algorithms combine their advantages.

In order to address the issue at hand and assess the effectiveness of the recommendation system, this article utilizes data sourced from the Google Local Dataset [6], with a specific focus on the geographical region of Alaska. To ensure the efficiency of the training dataset, an initial data preprocessing phase was conducted, which involved the removal of a substantial volume of non-informative data entries lacking specific evaluations and relevant information.

Subsequently, the algorithm part of this project mainly adopts hybrid methods, Initially, Singular Value Decomposition (SVD) technology [7] is employed to impute missing data values. Concurrently, a K-means clustering method [8] is utilized to process and categorize textual components.

The functionalities incorporated encompass two key aspects: cold start rating recommendations and personalized recommendations derived from customers' historical purchase behavior.

To facilitate the experimentation and assessment of the recommendation system, the dataset is partitioned into a sample set and a training set. Subsequently, the purchase rate is evaluated by aggregating results from all test samples.

2 Method

2.1 Dataset preparation

Based on one of the largest datasets currently available, the Google Online Shopping Dataset [6], Alaska's product selection was chosen as a case study due to the large amount of data. This dataset includes 427, 808 pieces of data, but a significant portion of these entries lack comments or contain concealed personal information, making the dataset relatively inconsistent and unsuitable for direct processing.

According to the feedback results of the dataset, the evaluation information of the product mainly comes from four aspects: 1. The user's rating after purchase 2. The user's evaluation 3. Location information 4. Related business information.

Due to the many flaws in the directly obtained dataset, the preprocessing part is essential. The first step is to perform text cleaning, which consists of three steps: 1. Remove special characters and punctuation; 2. Convert to lowercase; 3. Remove stop words. These tasks can effectively reduce noise, computation time, and case differences.

Later on, word splitting and blank filling are required for the text section. Based on this, this project also divided the dataset and randomly selected some personnel from the dataset to form a test set, and used the remaining ones as training sets to ensure the independence of the test samples. Moreover, since all datasets come from a unified region, the recommendation accuracy is higher, which is in line with reality.

2.2. Model

In a comprehensive recommendation system, model training is a key step that involves combining the results of Singular Value Decomposition (SVD) and K-means clustering to generate the final recommendation model. The following are specific explanations for each part SVD [7]

Using SVD technology to decompose the user item interaction matrix. SVD decomposes the original matrix into the product of three matrices [9]:
U-matrix (user matrix): Contains the relationship between users and potential features.

singular value matrix (Σ Matrix): Contains singular values that describe the importance of potential features. V ^ T matrix (transposition of item matrix): Contains the relationship between items and potential features.

Using decomposed U Σ Train the SVD model using the V ^ T matrix. The goal of training is to minimize the root mean square error (RMSE) or other loss functions between the original user item interaction matrix and the approximation of SVD decomposition.

The K clustering recommendation system uses K-means clustering to divide all words into multiple clusters (groups), each containing words with similar meanings [10]. This can facilitate the generation of varied recommendations, as words within distinct clusters may be pertinent to various user profiles, and there may also be shared words among different user types. Once the K-means clustering of items is completed, the recommendation system can select clusters that users may like based on their historical behavior and preferences.

The recommendation system can randomly select items from the selected cluster or select representative items in the cluster to recommend them to users.

By utilizing the comprehensive features constructed by SVD and K-means clustering, a comprehensive recommendation model with intelligent capabilities that can deeply understand and fuse these features was cultivated, thereby effectively generating personalized recommendation suggestions for users. This model combines the potential correlation of SVD with the diversity of K-means clustering to provide each user with a more accurate and diverse recommendation experience, ensuring that the final recommendation results meet the user's interests and needs.

2.3 Function

Cold start (without shopping data). Based on the feature that the recommendation system can start cold without any external input, this project automatically provides the best top 10 recommended products based on the entire data system when users first use it, and the proportion of address information in the recommendation system has significantly increased, allowing customers to complete their first transaction smoothly and confidently.

Personalized recommendations (Based on past data). After the user makes their first purchase, this system can make more accurate recommendations based on their feedback information. It analyzes user needs through clustering of similar items and evaluations, and classifies them to provide personalized recommendations that are suitable for the user. In terms of usage, only the user and the last purchased product need to be entered for analysis, without the need for additional data.

2.4 Performance testing

At the initial stage of the project, the dataset is segmented, with 10% being used as the test set and the rest as the training set. This method can ensure the randomness and authenticity of test samples. According to the defined evaluation indicators, the project continuously adjusts the parameters between each project, thereby improving the final purchase indicators.

3 Results and discussion

The Rating scoring system is an important part of collaborative filtering, and Fig 1 displays the comprehensive statistics of the top 15 products. Through statistics on the frequency of occurrence, it can be found that some items, although rated relatively low, have a high frequency of occurrence, indicating that people have a high demand for daily necessities and need to purchase them repeatedly. Therefore, the recommendation index for this product is actually better, and it is also in line with a wider consumer group.

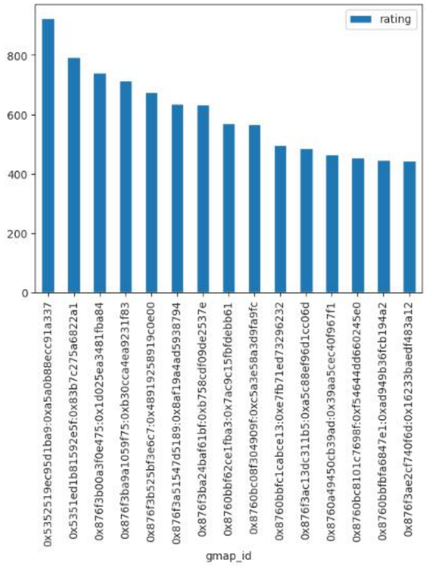


Fig. 1. Perform statistical analysis on the given rating (based on purchase frequency and rating) (Photo/Picture credit : Original).

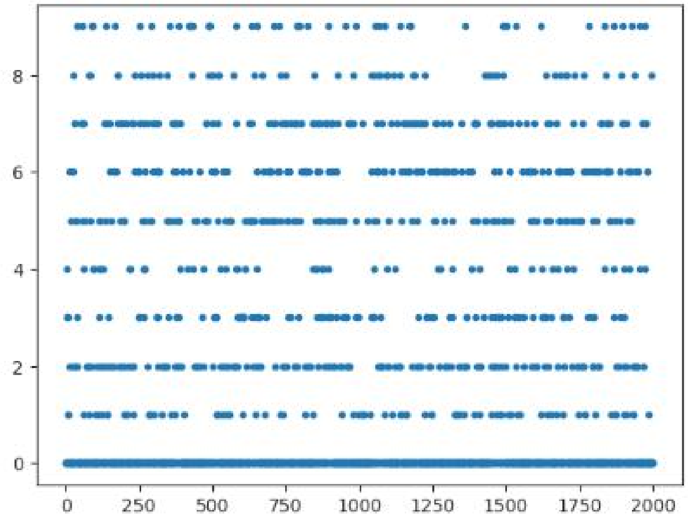
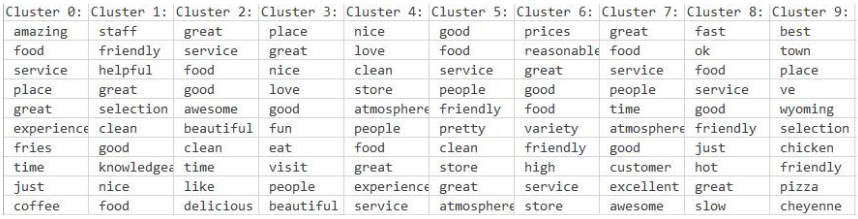
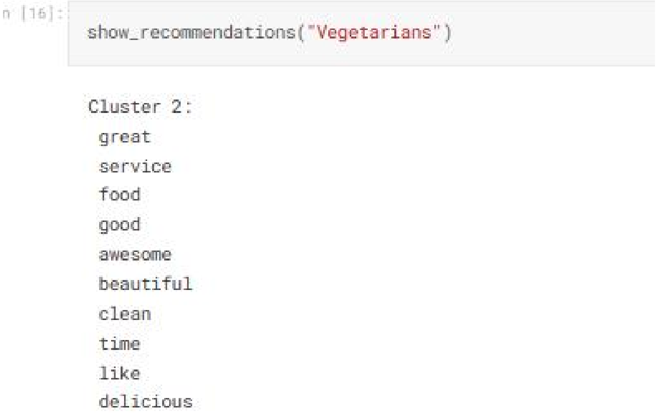


Fig. 2. The results of clustering and partitioning the divided words (Photo/Picture credit : Original).



Cluster 0:	Cluster 1:	Cluster 2:	Cluster 3:	Cluster 4:	Cluster 5:	Cluster 6:	Cluster 7:	Cluster 8:	Cluster 9:
amazing	staff	great	place	nice	good	prices	great	fast	best
food	friendly	service	great	love	food	reasonable	food	ok	town
service	helpful	food	nice	clean	service	great	service	food	place
place	great	good	love	store	people	good	people	service	ve
great	selection	awesome	good	atmosphere	friendly	food	time	good	wyoming
experience	clean	beautiful	fun	people	pretty	variety	atmosphere	friendly	selection
fries	good	clean	eat	food	clean	friendly	good	just	chicken
time	knowledgea	time	visit	great	store	high	customer	hot	friendly
just	nice	like	people	experience	great	service	excellent	great	pizza
coffee	food	delicious	beautiful	service	atmosphere	store	awesome	slow	cheyenne

Fig. 3. Display of key words in clustering results (Photo/Picture credit : Original).

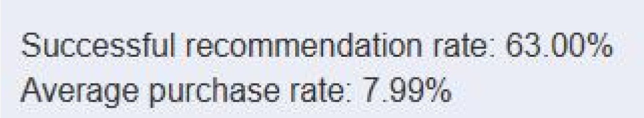


```
n [16]:  show_recommendations("Vegetarians")

         Cluster 2:
          great
          service
          food
          good
          awesome
          beautiful
          clean
          time
          like
          delicious
```

Fig. 4. Testing of clustering results (Photo/Picture credit : Original).

According to the clustering analysis results (see Fig 2), most words are classified into the first category, which is Cluster 0. This conclusion indicates that many people's online shopping habits are similar. And through the results (see Fig 3), it was found that different categories have the same words, indicating that people will also have similar evaluations of different things. Fig. 4 also presents the testing of clustering results. Therefore, the recommendation system cannot determine the category based on a single word and should judge it by all words.



```
Successful recommendation rate: 63.00%
Average purchase rate: 7.99%
```

Fig. 5. Display of comprehensive recommendation system results (Photo/Picture credit : Original).

Perform statistics on all data in the test set to obtain results (Fig 5). In fact, the comprehensive recommendation system is successful because 63% of people's choices are within the recommendation system, but the average purchase rate is only 7%, indicating that not all products have been selected.

By analyzing several recommendation results, it was found that most of them were similar products. This leads people to only choose one purchase from similar products due to overlapping features. This problem cannot be solved in datasets that do not classify products, so hope that future product datasets will include this.

4 Conclusion

This article builds a thorough recommendation system using collaborative filtering techniques, addressing the need for both cold start and personalized recommendations in contemporary e-commerce platforms. Nonetheless, this approach is not without its challenges. An analysis of the test set's final results reveals a high repurchase rate of 63.00%, indicating the potential for personalized recommendations for the user base. However, the actual outcomes have not met expectations, with an average user purchase index of only 7.99%. This is because the recommended samples are all similar to the same type of products, but there is no product type classification in the existing dataset, this should be noted in subsequent data collection.

In addition, the model also has room for improvement, and the dataset is not fully utilized. For this comprehensive recommendation algorithm, it is necessary to collect as many parameters as possible to refine the model. In the future, through comprehensive models, more detailed and accurate classification of purchasing population will be carried out, and real-time recommendations will be made based on practical factors.

References

1. Isinkaye, F. O., Folajimi, Y. O., Ojokoh, B. A.: Recommendation systems: Principles, methods and evaluation. Egyptian informatics journal, 16(3): 261-273 (2015).
2. Su, X., Khoshgoftaar, T. M.: A survey of collaborative filtering techniques. Advances in artificial intelligence (2009).
3. Bennett, J., Lanning, S.: The netflix prize. Proceedings of KDD cup and workshop 35 (2007).
4. Linden, G., Smith, B., York, J.: Amazon. com recommendations: Item-to-item collaborative filtering. IEEE Internet computing, 7(1): 76-80 (2003).
5. Grbovic, M., Radosavljevic, V., Djuric, N., et al.: E-commerce in your inbox: Product recommendations at scale. Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, 1809-1818 (2015).
6. Google Local Data. https://datarepo.eng.ucsd.edu/mcauley_group/gdrive/googlelocal (2021)
7. Henry, E. R., Hofrichter, J.: Singular value decomposition: Application to analysis of experimental data. Methods in enzymology. Academic Press, 210: 129-192 (1992).
8. Hartigan, J. A., Wong, M. A.: Algorithm AS 136: A k-means clustering algorithm. Journal of the royal (1979).

9. Hoecker, A., and Kartvelishvili, V.: SVD approach to data unfolding. Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, 372(3), 469-481 (1996).

10. Qiu, Y., Chen, P., Lin, Z., et al.: Clustering Analysis for Silent Telecom Customers Based on K-means++, 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC). IEEE, 1: 1023-1027 (2020).