Proceedings of the 5th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/15/20230813

Research on accuracy analysis and improvement of recommender system based on amazon review

Xuyang Wang

School of Electronic and Information Engineering, Lanzhou Jiaotong University, Lanzhou, 730070, China

20213208134@stu.lzjtu.edu.cn

Abstract. Recommender system is a system that uses artificial intelligence and data mining technology to recommend items or services that meet users' preferences based on their historical behaviors and interests. In modern society, people are faced with more and more choices, and the emergence of recommender system can help users filter out valuable content from complex information and improve user satisfaction and experience. In this paper, collaborative filtering is used to implement a recommender system based on Amazon review data set. Meanwhile, Singular Value Decomposition (SVD) and Principal Component Analysis (PCA) are used to conduct dimensionality reduction and other operations on the data. Root Mean Squared Error (RMSE) A value expressed as a recommendation for accuracy. After the establishment of the recommender system and three precision analysis experiments, it achieves this by applying a selected filtering algorithm to the input supplied, which is frequently in the form of user reviews of products.

Keywords: recommender systems, collaborative filtering, singular value decomposition (SVD), principal component analysis (PCA).

1. Introduction

The recommender system is designed to simulate sales staff to inform clients about products and make product recommendations in order to assist users in making purchasing decisions. In today's Internet age, e-commerce, social media, music, movies and other fields are developing rapidly, and a wide variety of various products have emerged. How to recommend information and products that users are interested in from massive data according to the user's interest characteristics and purchase behavior contains huge commercial value. In recent years, the study of recommender systems is particularly interesting to both academics and business.

In recommender systems, there are two basic entities: users and items. Users of recommender systems are required to comment on earlier items. A recommender system's objective is to produce suggestions for brand-new products for these particular users. It accomplishes this by using a chosen filtering algorithm on the input given, typically in the form of user ratings on items [1]. The problem of recommending items from a database has received a great deal of attention, and two main paradigms have emerged. The three primary methods for making suggestions are content-based filtering, collaborative filtering, and hybrid techniques. A movie's genre, director, and star are just a few examples of discrete aspects of what a content-based filtering algorithm can use to generate recommendations.

© 2024 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/15/20230813

Collaborative filtering suggestions aim to provide a list of desirable items for engaged users based on the preferences of their like-minded group [2-4]. In content-based recommendation, although in collaborative suggestion, people with similar likes to a given user are found and products they enjoy are recommended, items comparable to items that a specific user has previously liked are recommended. [5]. In order to create hybrid recommendations systems, these two approaches are often used in combination [2-4].

The most popular method of suggestion is collaborative filtering [6-7]. There are several methods for content-based analysis, including spectrum analysis, latent semantic models, matrix factorization, and social recommendation [8–13]. All of these strategies advise products that users who are comparable to the target user for whom the suggestions are being computed have enjoyed (similarly, item-based approach builds on evaluating the similarity of items). Because to their significant role in the winning solution for the Netflix reward competition, the last listed class of algorithms has lately acquired notoriety. [14-16]. Based on the requirements of the above data files and recommender system, collaborative filtering recommendation meets the requirements of the above tasks.The recommender system uses the collaborative filtering algorithm to generate suggestions, which may locate other users who have a high degree of similarity to users based on previous data and recommend their preferred things.

By analyzing users' historical data and interests, the recommender system can improve users' shopping experience, reduce their selection difficulties, and increase their adherence to the platform. The recommender system can simultaneously help retailers boost sales, lower inventory, improve pricing, etc. On the basis of the data files of Amazon sports outdoor reviews from 2014 to 2018, a recommender system will be built and recommend appropriate products to users based on the goods users have previously purchased, rated, and the purchase scores of users who are similar to them. Experimental analysis is carried out to further understand the recommender system.

In this paper, focusing on the above-mentioned aspects, the following section 2 Recommender System basic information will describe the basic information about the recommender system, including data set examples, basic concepts related to the recommender system, the model and algorithm used by the recommender system, as well as specific details of the design and operation of the recommender system. The section 3 Experiment and analysis of accuracy of recommender system is about three experiments to test the recommendation accuracy of the system, with RMSE as the evaluation index. Discuss problems and summarize options ask how the recommender system should improve the accuracy, i.e., reduce the value of RMSE, and list the corresponding methods to reduce the value of RMSE. The section 5 Conclusions is the output of the whole paper and the experimental results.

2. Recommender system basic information

2.1. datasets

The 2014-released Amazon review dataset has been updated using this dataset. Links, reviews (ratings, text, helpfulness votes), product metadata (descriptions, category information, price, brand, and picture properties), and also viewed/also bought graphs are all included. The details of original data set are illustrated in following Table 1 and Table 2. In the data set, the meaning of different attributes are as follows in Table 3.

2.2. Basic principles theoretical concepts

A number of fundamental issues plague recommender systems, reducing the accuracy of forecasts made. Examples of these problems include synonymy, sparsity, and scalability. To deal with this, several alternatives have been put up. It is particularly interested in mathematical procedures that successfully address the aforementioned problems by identifying efficient methodsto make the starting data less dimensional. Singular Value Decomposition (SVD) and Principal Component Analysis (PCA) are two examples of such methods [1].

DOI: 10.54254/2755-2721/15/20230813

Table 1. Samples of the review data of the original data set.

overall	verified	reviewTime	reviewerID	asin	reviewerName
5	TRUE	06 3, 2015	A180LQZBUWVOLF	32034	Michelle A
1	TRUE	04 1, 2015	ATMFGKU5SVEYY	32034	Crystal R
5	TRUE	01 13, 2015	A1QE70QBJ8U6ZG	32034	darla Landreth
5	TRUE	12 23, 2014	A22CP6Z73MZTYU	32034	L. Huynh
4	TRUE	12 15, 2014	A22L28G8NRNLLN	32034	McKenna

Table 2. Samples of the reviewText of the original data set.

reviewText	summary	unixReviewTime	style	vote	image
What a spectacular tutu!	Five Stars	1433289600	NaN	NaN	NaN
What the heck? Is this ...	Is this a tutu for nuns?	1427846400	NaN	NaN	NaN
Exactly what we ...	Five Stars	1421107200	NaN	NaN	NaN
I used this skirt for ...	I liked that the elastic ...	1419292800	NaN	NaN	NaN
This is thick ...	This is thick enough ..	1418601600	NaN	NaN	NaN

Table 3. The meaning of different attributes.

Title	Meaning
overall	rating of the product
reviewTime	time of the review (raw)
reviewerID	ID of the reviewer, e.g. A2SUAM1J3GNN3B
asin	ID of the product, e.g., 0000013714
reviewerName	name of the reviewer
reviewText	text of the review
summary	summary of the review
unixReviewTime	time of the review (unix time)
style	a disctionary of the product metadata, e.g., "Format" is "Hardcover"
vote	helpful votes of the review
image	images that users post after they have received the product

2.2.1. Principal component analysis (PCA). Principal component analysis (PCA), a statistical technique, is used to reduce the dimensionality of datasets while maintaining the integrity of the important data. It is a multivariate mathematical method that takes a collection of variables that might be connected and turns them into a new set of uncorrelated variables. The starting variables are combined linearly to form its principal components, which are its constituent parts. Usually, the variables are organized in decreasing order of degree of variation, with the first principal component including the variables with the highest degree of variation and each subsequent principal component reflecting the next highest degree of variation [17]. Data compression, visualization, and feature extraction are just a few of the many uses for PCA in data analysis and machine learning. By reducing the number of variables in a dataset, PCA can speed up processing and improve the efficacy of machine learning algorithms, especially when working with high-dimensional datasets.

2.2.2. Singular value decomposition (SVD). Singular Value Decomposition (SVD) is a common linear algebra technique, which is mainly used to decompose and reduce dimensionality of high-dimensional matrix [18]. It breaks down a matrix into the sum of three other matrices. In some cases, the true meaning

DOI: 10.54254/2755-2721/15/20230813

behind a data matrix cannot be found, but the essential information in the data matrix is acquired by singular value decomposition. SVD are often used in the fields of data dimension reduction, data compression, matrix approximation, etc. The following are the benefits of using SVD in a collaborative filtering recommendation algorithm: Because SVD can uncover the underlying characteristics behind the data matrix, they enable users to establish a closer relationship with the project, which significantly improves the precision of the recommended outcomes. In this recommender system, SVD are used to decompose the user-item scoring matrix, so as to find the potential user and item characteristics in the low-dimensional space, so as to realize the recommendation.

2.3. Details of model construction

2.3.1. Data processing. The input data is pre-processed to calculate the average rating for each unique combination by grouping users, items, and time. After then, only users who had rated at least five things in the datasets and had done so in 2018 or later were allowed to access the data. The processed data is then transformed into Surprise datasets objects, as shown in Table 4.

Table 4. Samples of processed data.

asin	reviewerName	reviewTime	AVGoverall
0000032034	Crystal R	04 1, 2015	3.571429
0000032034	JmeEd	02 7, 2016	3.571429
0000032034	L. Huynh	12 23, 2014	3.571429
0000032034	McKenna	12 15, 2014	3.571429
0000032034	Michelle A	06 3, 2015	3.571429
...	...	...	...
B01HJHHBHG	medinaroger	04 27, 2017	5.000000
B01HJHHBHG	PJT	10 28, 2017	5.000000
B01HJHHBHG	Steve	02 13, 2017	5.000000
B01HJHHBHG	goosedowner	06 11, 2017	5.000000
B01HJHHBHG	old hunter	03 17, 2018	5.000000

2.3.2. Model training. A training set and a test set are created from the data set, and an SVD algorithm with 100 potential factors is trained using the training set. Then, PCA was used to reduce. To speed up the computation of suggestions, the user and item dimensions are multiplied by 50.

2.3.3. Recommendation. Define a "recommend" function that takes the user's name as input and returns the user's top 5 recommended items. It first converts the username into the internal ID used by Surprise library, obtains the user's dimensionality reduction feature vector from the trained SVD algorithm, calculates the similarity score between the user and all items in the dimensionality reduction feature space using dot product, and chooses the best five recommendations based on their score. Finally, use the Surprise library's utility function to convert the final recommended item back to its original item ID.

2.3.4. User input and output. Prompts the user for his or her username and checks to see if the user is in the training set. If the user is in the training set, call the "recommend" function to generate a personalized recommendation and print it out. Otherwise, it prints out an error message. As shown in Table 5, we give some examples of the product recommended.

DOI: 10.54254/2755-2721/15/20230813

Table 5. Examples of the product recommended.

asin	userName	
	Andrew M. Silverman	Blake Zimmerman
asin1	7245456313	BO00051ZHS
asin2	B0004TBLW	B0004TBLW
asin3	B000051ZHS	B00004U31L
asin4	B00004NKIQ	B00004T11T
asin5	B00002N6T4	7245456313

3. Experiment and accuracy analysis

3.1. Evaluation metric

The root mean squared error calculates the discrepancy between predicted and actual numbers (RMSE). In the context of machine learning, it is frequently employed, particularly when evaluating neural networks. By calculating the square root of the mean squared errors, RMSE is obtained. The squared error is the difference between the actual value and the anticipated value. For expected values, RMSE is a useful metric of accuracy. The test set's root mean square error (RMSE) is calculated by using the "accuracy" module of Surprise library to assess the trained algorithm's performance. Overall, the recommender system uses the Surprise library to pre-process, train, and evaluate the recommender system, as well as how to generate personalized recommendations based on past user ratings.

3.2. Loss curve in model training

In the experiment, to study the loss of the model in training, The training set's and test set's loss curves are examined., which is shown in Figure 1. This loss curve reveals that there is an overfitting issue with the model because the RMSE values of the training set and test set show no discernible changes. Overfitting is the term for a model's performance when it is good on the training set but poor on the test set. When a model is overfit, it learns the specifics and noise of the training data while neglecting its general characteristics and capacity for generalization. Therefore, on the test set, the model may not properly generalize to the new data.

Specifically, when the loss curves of the training set and the test set are parallel but greatly different, it often denotes an overfitting of the model to the training set. Alternatively, the model overmatches the data in the training set, resulting in a very tiny error on the training set, but when generalized to the test set, the error on the test set is quite significant, meaning that it is considerably different from the error on the training set.

To solve this problem, regularization techniques can be used to limit the model's complexity and avoid overfit. In addition, more data can be added, or data enhancement techniques can be used to broaden the range of the data and enhance the model's generalizability.

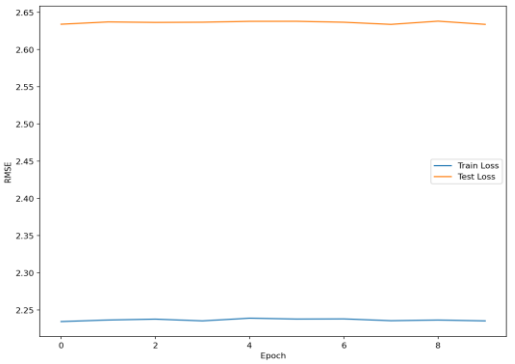

Figure 1. Loss curve of the training and test set.

DOI: 10.54254/2755-2721/15/20230813

3.3. Parameters analysis

In this experiment, different combinations of SVD and PCA parameters were selected to run the effectiveness of the recommender system and the recommender system was calculated. RMSE is used as an evaluation index to represent the root mean square error of the recommender system.Table 6 are the results of RMSE changes that change PCA parameters, SVD parameters, etc. In this table, it can be found that for this model, The recommender system's accuracy of recommendations is unaffected by a change in PCA settings. The suggestion accuracy varies clearly when the SVD value is modified. The system's suggestion accuracy can be increased by the decrease of SVD parameters.

Table 6. Change of RMSE with different parameters settings.

SVD	PCA									
	10	20	30	40	50	60	70	80	90	100
10	2.4763	null	null	null	null	null	null	null	null	null
20	2.5121	2.5059	null	null	null	null	null	null	null	null
30	2.5324	2.5337	2.5316	null	null	null	null	null	null	null
40	2.5599	2.553	2.5606	2.5485	null	null	null	null	null	null
50	2.5688	2.5774	2.5759	2.5801	2.577	null	null	null	null	null
60	2.5854	2.5856	2.5902	2.5977	2.5915	2.5953	null	null	null	null
70	2.5989	2.6059	2.5979	2.6069	2.6076	2.6087	2.609	null	null	null
80	2.6163	2.621	2.6106	2.6121	2.6113	2.6127	2.6176	2.6182	null	null
90	2.6206	2.6285	2.6296	2.6149	2.6372	2.6294	2.6284	2.6269	2.6295	null
100	2.6342	2.6409	2.644	2.6415	2.6363	2.6391	2.6455	2.6401	2.6499	2.6451

3.4. Performance for different numbers of recommended products

It is a difficult problem to determine how the quantity of items in the recommender system affects accuracy, because it depends on many factors, such as data set, recommendation algorithm, user characteristics and so on. Generally speaking, the more items are recommended in the recommender system, Users may find it simpler to discover the products they want, but there are certain drawbacks, such as: (1) Too long a recommendation list will increase the cost of user selection, and users need to spend more time and energy to browse the recommendation list. (2) Because it might be challenging for users to locate the products they are truly interested in, extensive lists of recommendations can reduce the effectiveness of recommendations. (3) The accuracy of suggestions may be compromised by excluding certain things that the user may find interesting in a list that is too short. As a result, in real-world applications, it is necessary to consider various factors and choose the appropriate quantity of recommended products for the recommender system.

Table 7. The RMSE of model when the recommended products increasing from 1 to 95.

1	5	10	15	20	25	30	35	40	45
2.6457	2.6456	2.6410	2.6260	2.6458	2.6377	2.6439	2.6317	2.6396	2.6367
50	55	60	65	70	75	80	85	90	95
2.6429	2.6341	2.6384	2.6421	2.6416	2.6441	2.6460	2.6308	2.6403	2.6394

The RMSE value in this recommender system is unaffected whether the experimental findings provided in Table 7 above indicate that the recommended number of elements is between 5, 10 to 95. If the number of suggested products is altered, the RMSE value is not affected, possibly for the following reasons: (1) Sparsity of the data set. If the data set used is very sparse, increasing the number of recommended items may not have a significant impact on the RMSE. This is because even if more recommended items are added, the user is likely to have no rating record, causing the RMSE value to remain the same. (2) The algorithm's performance in making recommendations. Increasing the number of recommended items likely has no appreciable impact on the RMSE because the recommendation system performs well enough. In this case, it may be necessary to use other metrics or evaluation

DOI: 10.54254/2755-2721/15/20230813

methods to more fully evaluate the performance of the recommendation algorithm. (3) Problems with experimental design. There may be problems with experimental design, such as insufficient changes in the number of recommended items to significantly affect RMSE, or problems with experimental data. The specific cause of the problem can be determined by a more detailed analysis of the experimental design.

Through the analysis of the performance and data set of the recommender system, it can be concluded that the reason why RMSE is not affected when the number of recommended items is changed is Article 2 above: The performance of the recommendation algorithm is probably good enough that increasing the number of recommended items has no significant effect on the RMSE.

4. Discuss problems and summarize options

There are still some problems in this recommender system. For example, the value of RMSE is too large, and the prediction error of the recommender system is relatively large. No cross-validation is used to more precisely validate the recommender system. Here are some ways to lower RMSE:

(1) Use additional data. The model's ability to accurately represent the distribution of scores is enhanced as the size of the data set grows.

(2) Addition of features. The model's accuracy and error may be increased by adding new features, such as user history score, commodity category, time, etc.

(3) Make changes to the model's parameters. To improve the model's performance in the collaborative filtering process, for instance, the hidden vector dimension, regularization parameters, learning rate, and other super parameters can be changed.

(4) Use an integrated approach. Combining multiple models can reduce prediction errors, such as using random forests, gradient lifting trees, etc.

(5) Cross-validation. You may assess your model's performance and choose the ideal set of model parameters and features by using cross-validation approaches.

5. Conclusions

This study first provides a thorough description of a recommender system based on Amazon review data. Three experiments are used to examine and verify the recommender system's accuracy in order to gain a greater knowledge of it: loss curve, changing key parameters and the number of recommended items. This work is an integral part of the design and implementation process of the recommender system. The experimental findings indicate that SVD parameters have an effect on the accuracy of the recommender system, and its operation is satisfactory.

References

[1] Vozalis M.G., Margaritis K.G. “A Recommender System using Principal Component Analysis”, published in 11th panhellenic conference in informatics, 2007, pp. 271-283.

[2] Bobadilla J., et al., “Recommender systems survey”, Knowl. -based Syst 46, 2013, pp. 109-132.

[3] Dietmar Jannach,et al., “Recommender Systems: An Introduction”, Cambridge University Press, USA, 2010.

[4] Feng Zhang, et al., “Fast algorithms to evaluate collaborative filtering recommender systems”, Knowl. -based Syst 96, 2016, pp. 96-103.

[5] Balabanović M. and Shoham Y. “Fab: content-based, collaborative recommendation”, Commun. ACM, vol 40, no 3, pp. 66-72, 1997.

[6] Goldberg, D, et al., “Using collaborative filtering to weave an information tapestry”, Commun. ACM, vol 35, pp. 61-70, 1992.

[7] Schafe J.B., et al., “Collaborative filtering recommender systems”, The Adaptive Web, Springer, 2007, pp. 291-324.

[8] M.J. Pazzani and D. Billsus. “Content-based recommender systems”, The Adaptive Web, Springer, 2007, pp. 325-341.

[9] K. Goldberg, et al., “Eigentaste: A Constant Time Collaborative Filtering Algorithm”, Inf.

DOI: 10.54254/2755-2721/15/20230813

Retr., vol 4, no 2, pp. 133-151, 2001.
[10] T. Hofmann. “Latent semantic models for collaborative filtering”, ACM Trans. Inf. Syst., vol 22, pp. 89-115, 2004.
[11] Y. Koren, R. Bell and C. Volinsky. “Matrix factorization techniques for recommender systems”, Computer, vol 42, pp. 30-37, 2009.
[12] U. Shardanand and P. Maes. “Social information filtering: algorithms for automating ‘word of mouth’”, Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM Press/Addison-Wesley Publishing Co. pp. 210-217, 1995.
[13] H. Ma, et al., “Recommender systems with social regularization”, Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, ACM, pp. 287-296, 2011.
[14] Y. Koren. “Collaborative filtering with temporal dynamics”, Commun. ACM, vol 53, pp. 89-97, 2010.
[15] J. Bennett and S. Lanning The netflix prize, Proceedings of KDD Cup and Workshop, vol 9, p. 35, 2007.
[16] Yu F., et al., “Network-based recommendation algorithms: A review”, Physica A, vol 452, pp. 192-208, 2016.
[17] Jolliffe T.I. Principal Component Analysis, Springer, New York, 2002.
[18] Ying Zhao, “Collaborative Filtering Algorithm Based on SVD”, Communications World, vol 293, no 10, p. 255, 2016.