Intelligent recommendation of related items based on naive bayes and collaborative filtering combination model

WeiWei, ZhuoWang, Changlong Fu, Robertas Damaševičius, Rafał Scherer, and Marcin Woźniak

School of Computer Science and Engineering, Xi'an University of Technology, Xi'an 710048; Shaanxi Key Laboratory for Network Computing and Security Technology, China.
E-mail: weiwei@xaut.edu.cn, clfu@xaut.edu.cn.

Abstract. Nowadays, data plays a unique role in various fields. Based on the background of the era of big data, this paper collects user evaluations of certain commodities and labels the evaluations into positive emotions and negative emotions. We propose an intelligent recommendation algorithm based on Naive Bayes algorithm, which determines the user's preference for a product according to the user's opinion on a product, then uses the collaborative filtering algorithm to recommend other similar products according to the similarity between the product and other products, so as to realize the recommendation of other similar products. This can not only help users choose more goods in the environment, but also maximize the value of goods. In this paper, bayesian classifier was used for prediction, and the accuracy reached about 97%.

1. Introduction

In today's era of big data [1], the information seems to have been the phenomenon of overload, and the slowly recommendation system algorithm [2-4] ability in the field of electronic commerce gradually revealed. In a certain extent, optimize the electronic commerce is the advantage of real-time intelligent recommendation system for users. It can accurately recommend the historical information of commodities, maximize the benefits of e-commerce, improve the accuracy of e-commerce system, and save shopping time for users. There are still many problems in real time and other aspects. Based on the existing problems, this paper integrates the model of bayesian classifier and collaborative filtering algorithm, and then predicts how much emotion users have for a certain type of goods, and recommends similar goods based on the emotion degree.

In recent years, with the continuous innovation of computer technology and the continuous and rapid development of network science, the recommendation system began to receive more attention from many scholars in all aspects. Therefore, in the actual network environment, information overload is often encountered. In order to alleviate this phenomenon, the recommendation system can make good use of it. Among many recommended algorithms, the collaborative filtering algorithm is one of the more classic and most used algorithms in this field. When users browse information on the Internet, they will leave a history or score of visits. The collaborative filtering algorithm will use this similarity to measure the similarity between users or items, determines the nearest neighbor set of the user or item, and generates corresponding recommendation services for different objects.

Collaborative intelligent recommendation is a method of recommending personalized recommendations based on purchases through collaborative filtering algorithms. Collaborative
intelligent recommendation includes user collaborative filtering and item collaborative filtering. And implement recommendations like users.

2. Model introduction

2.1. Bayesian classifier

Bayesian classification is the general name of a classification algorithm [5-6]. This algorithm is based on Bayes’ theorem, so it is called Bayesian classifier. The principle of the Bayesian classifier is to use the Bayesian formula to calculate the posterior probability of an object, that is, the object belongs to a certain probability, and selects the class with the largest posterior probability as the object class.

Set sample \( \mathbf{x} = (x^1, x^2, x^3, ..., x^n) \in X \) and set label \( y \in Y = \{c_1, c_2, c_3, ..., c_K\} \). Let \( X \) be the random vector above \( X \), \( Y \) be the random variable above \( Y \), and \( P (X, Y) \) be the joint probability distribution of \( X \) and \( Y \). Assuming that the training data set \( T = \{ (x_1, y_1), (x_2, y_2), (x_3, y_3), ..., (x_n, y_n) \} \) is independently and identically distributed by \( P (X, Y) \), then the Naive Bayes method can learn the joint distribution probability distribution \( P (X, Y) \) in the training data set, which is the following probability distribution map.

(1) Prior probability distribution:

\[ P(Y = c_k), k = 1, 2, 3, 4, ..., K \]

(2) Conditional probability distribution:

\[ P(X = x | Y = c_k), k = 1, 2, 3, 4, ..., K \]

Where \( c_k \) indicates the category to which \( Y \) belongs, and \( K \) indicates that the data is divided into several categories.

2.2. The collaborative filtering algorithm

The collaborative filtering algorithm relies on the user's historical behavior data [7-17], and does not need to obtain the user or item's characteristic data in advance. Therefore, the user's historical behavior data is used for modeling to achieve the purpose of user recommendation. Collaborative filtering is a method used to analyze user interests. By synthesizing these users' evaluation of certain information, a system is formed to predict the user's preference for a certain type of goods[10-19].

The key of the collaborative filtering algorithm is to calculate the similarity between users. Currently, similar calculation methods include Euclidean distance, Cosine similarity, and Tanimoto coefficient. This article uses the Euclidean distance, The common similarity calculation formulas are as follows. Similarity calculation is the key to finding users in the neighborhood. Similarity calculation mainly includes cosine similarity, modified cosine similarity, Pearson correlation coefficient, various distance similarities, etc[18-26].

(1) Cosine similarity

The cosine similarity is mainly determined by calculating the angle between two vectors, which is also called the angle cosine. The value is between -1 and 1. The larger the angle cosine, the greater the angle between the two vectors. The smaller, the higher the similarity. The specific calculation formula is show in (1) and (2):

(2) Modified cosine similarity

\[ d(x, y) = \sqrt{\left(\sum(x_i - y_i)^2\right)} \]  

\[ \text{sim}(x, y) = \frac{1}{1 + d(x, y)} \]  

(3) Pearson correlation coefficient

The Pearson correlation coefficient, also known as correlation similarity, needs to find items that two users have rated together and then calculate their correlation [25]. The calculation formula is show in (3):
Distance similarity

The cosine similarity introduced above mainly focuses on the difference degree of the vector direction, while the distance similarity focuses on the distance between two points, the closer the distance, the greater the similarity. Distances include Euclidean distance, Manhattan distance, Chebyshev distance, Minkowski distance, Mahalanobis distance, etc., because distance and similarity are roughly inversely proportional. The distance similarity formula is show in (4):

\[ \text{sim}(u, v) = \frac{\sum_{i \in I_u, j \in I_v} (r_{ui} - r_u)(r_{vi} - r_v)}{\sqrt{\sum_{i \in I_u} (r_{ui} - r_u)^2} \sqrt{\sum_{j \in I_v} (r_{vj} - r_v)^2}} \]  

(3)

\[ \text{sim}(u, v) = \frac{1}{1 + d(u, v)} \]  

(4)

Common score prediction methods

Average scoring

The average scoring method is based on the average value of all the scores of the neighboring users on the target item as the target user's score on the target item [26]. Suppose the neighbor user is \( U = (u_1, u_2, \ldots, u_m) \) and the item is \( I = (I_1, I_2, \ldots, I_n) \). The specific formula is show in (5):

\[ r(u, i) = \frac{1}{n} \sum_{k \in U} r(k, i) \]  

(5)

Weighted average scoring method

The average scoring method takes the average of the user score data in all neighbors as the target score, but ignores the influence of the similarity weight. The higher the similarity with the target user, the more accurate the score prediction result, so the weighted average score method introduces similarity. The weight, \( s(u, k) \) is the similarity between the target user \( u \) and the neighbor user \( k \) [27], the specific formula is show in (6):

\[ r(u, i) = \frac{\sum_{k \in U} s(u, k)r(k, i)}{\sum_{k \in U} s(u, k)} \]  

(6)

3. Data acquisition and preprocessing

3.1. Datasources

This article uses crawler technology to crawl a certain shopping APP. In order to save article space, this article shows some data as shown in Table 1 below:

| No. | comments |
|-----|----------|
| 1   | On the screen, Honor Play4T Pro uses a 6.3-inch pearl screen, and the screen material of Honor Play4T Pro ... |
| 2   | Appearance: It is very comfortable and good to hold in hand, screen sound effect: very amazing, for this price point, the sound is also good, the photo effect: very ... |
| 3   | Appearance: very beautiful two-color back cover is very nice, screen sound effect: good sound quality, photo effect: Huawei's photo effect is very good, running ... |

3.2. Data pre-processing

(1) The stop words are removed to solve the useless high-frequency words, numbers and English letters in the text, which will improve the accuracy of the results. After removing the stop words, the text is segmented. This article uses the jieba library for processing, and then counts the word frequency.

(2) Since the captured data does not have emotional tags, you need to tag each comment. Here, according to the sentiment of each comment, the positive emotion tag is set to 1, while the negative
emotion tag is set to be unique. If it is 0, this article will manually label 600 pieces of data. As shown in Table 2 below:

|   | comments                                                                 | label |
|---|---------------------------------------------------------------------------|-------|
| 2 | On the screen, Honor Play4T Pro uses a 6.3-inch pearl screen, and the screen material of Honor Play4T Pro ... | 1     |
| 3 | Appearance: It is very comfortable and good to hold in hand, screen sound effect: very amazing, for this price point, the sound is also good, the photo effect: very ... | 1     |
| 4 | Appearance: very beautiful two-color back cover is very nice, screen sound effect: good sound quality, photo effect: Huawei's photo effect is very good, running ... | 1     |

Finally, we need to construct the word vector space of the comments. Among many natural language processing tasks, benefit from word vectors trained on large-scale corpora. Because pre-trained word vectors have general semantic features on large corpora, when these word vectors are applied to specific downstream tasks, they often need to be updated and adjusted through fine-tuning to make them more suitable for target task training. However, the low-frequency words in the target corpus often lack the training samples, which makes it impossible to obtain stable gradient information during the correction process, so that the word vector cannot be effectively updated. In the short text classification task, these low-frequency words are also important indicative of the classification results. Therefore, it is necessary to obtain a better representation of low-frequency words in specific short text classification tasks. Because the comment data in this experiment is short, there is a low-frequency word vector representation that can improve the accuracy of the experiment[19-26].

4. Analysis of results

For the credibility of this paper, the data predicted by the Bayesian classifier are visualized accordingly, and the item recommendation is obtained by using the collaborative filtering algorithm.

4.1. Data pre-processing

This article crawled a total of 800 pieces of data and labeled 600 pieces of data. The data is divided into a training set and a test set. The remaining 200 pieces of data are not labeled as test sets. A Bayesian classifier is used to build a model of 600 pieces of data, and 200 pieces of data are used for prediction and analysis. The prediction results are shown in the confusion matrix in Figure 1 below:

![Fig.1 Good similarity prediction results](image)

It can be seen from the above figure that the result based on the Bayesian classifier reaches nearly 97%, which very well illustrates the effect of the Bayesian classifier.
4.2. Collaborative filtering
This article extracts the comments of a user who bought an item, and then recommends another item to the user according to the feature similarity between the items. The result is shown in Figure 2:

As can be seen from Figure 2 above, Item 1 and Item 2 have the highest similarity, so when they are recommended to the same user, the higher the similarity, the greater the probability of recommendation to the user.

5. Conclusion
In this paper, a Bayesian classifier is used to predict and analyze the user's emotions, determine the user's sentiment to a certain type of product, and then calculate the item-to-item similarity according to the collaborative filtering algorithm to obtain the final recommendation effect, but this article based on the data crawled in the APP, it may not be comprehensive. In the follow-up work, more data will be perfected. This article considers user shopping data and user information as model feature data. This makes recommendations more accurate and effective.

Acknowledgments
This job is supported by the National key R&D Program of China under Grant NO. 2018YFB0203901 and the Key Research and Development Program of Shaanxi Province (No.2018ZDXM-GY-036) and Shaanxi Key Laboratory of Intelligent Processing for Big Energy Data (No.IPBED7).

References
[1] Li Benyue, Li Weirong, Pan Huafeng, Wang Hong, Wang Qi. On the influence of artificial intelligence on diagnosis of traditional Chinese medicine [J / OL]. World Science and Technology—Modernization of Traditional Chinese Medicine, 2020:1-5.
[2] Rahim Rashidi, Keyhan Khamforoosh, Amir Sheikhamadi. An analytic approach to separate users by introducing new combinations of initial centers of clustering [J]. Physica A: Statistical Mechanics and its Applications, 2020, 551.
[3] David M. Brown, Louise Camenzuli, Aaron D. Redman, Chris Hughes, Neil Wang, Eleni Vaiopoulou, David Saunders, Alex Villalobos, Susannah Linington. Is the Arrhenius-correction of biodegradation rates, as recommended through REACH guidance, fit for environmentally relevant conditions? An example from petroleum biodegradation in environmental systems [J]. Science of the Total Environment, 2020:732.
[4] Sonya Dasharathy, Folasade (Fola) P. May, Anthony Myint, Liu Yang, Harman K. Rahal, Vivy Tran, Philip A. Kozan, Sarina C. Lowe, Peter Y. Beah, Berkeley N. Limketkai, Jenny S. Sauk. 545 pneumococcal vaccination recommendation and completion rates among inflammatory bowel disease patients within a large academic health system (poster presentation) [J]. Gastroenterology, 2020, 158(6).
[5] Xiang Wei, Zhou Wenxing. Integrated pipeline corrosion growth modeling and reliability analysis using dynamic Bayesian network and parameter learning technique [J]. Structure and Infrastructure Engineering, 2020, 16(8).
[6] Yu Jiang, Pablo J. González. Bayesian Inversion of Wrapped Satellite Interferometric Phase to Estimate Fault and Volcano Surface Ground Deformation Models [J]. Journal of Geophysical
Research: Solid Earth, 2020, 125(5).

[7] Deng Ailin, Zhu Yangyong, Shi Bole. Collaborative filtering recommendation algorithm based on item score prediction [J]. Journal of Software, 2003,(9):1621-1628.

[8] Bin Zhou, Dawid Polap, and Marcin Wozniak. A regional adaptive variational PDE model for computed tomography image reconstruction, Pattern Recognition, 2019, 92: 64-81.

[9] Xia X, Marcin W, Fan X, Damasevicius R., Li Y. Multi-sink distributed power control algorithm for Cyber-physical-systems in coal mine tunnels. Computer Networks, 2019, 161: 210-219.

[10] Song, H., Li, W., Shen, P., & Vasilakos, A. Gradient-driven parking navigation using a continuous information potential field based on wireless sensor network. Information Sciences, 2017, 408(2): 100-114.

[11] Xu Q, Wang L, Hei XH, Shen P, Shi W, Shan L, GI/Geom/1 queue based on communication model for mesh networks. International Journal of Communication Systems. 2014, 27(11): 3013-29.

[12] Fan, X., Song, H., Fan, X., & Yang, J. Imperfect information dynamic stackelberg game based resource allocation using hidden Markov for cloud computing[J]. IEEE Transactions on Services Computing, 2016, 11(1): 78-89.

[13] Zhang Jiangshe, Song Houbing, Yan Wan. Big data analytics enabled by feature extraction based on partial independence[J]. Neurocomputing, 2018, 288(3): 3-10.

[14] Zhang Jiangshe, Wei Wei, Dawid Polap, Marcin Wozniak, Leon Kosmider, Robertas Damasevicius, A neuro-heuristic approach for recognition of lung diseases from X-ray images[J]. Expert Systems with Applications, 2019, 126: 218-232.

[15] Su, J., Song, H., Wang, H., & Fan, X. Cdma-based anti-collision algorithm for epc global c1 gen2 systems[J]. Telecommunication Systems, 2018, 67:1-9.

[16] Sun Z, Song H, Wang H, Fan X. Energy Balance-Based Steerable Arguments Coverage Method in WSNs[J]. 2018, 6: 33766-33773.

[17] Song H, Wang H, Fan X. Research and Simulation of Queue Management Algorithms in Ad Hoc Network under DDoS Attack. IEEE Access, 2017, 5: 27810-27817.

[18] Qiang Y, Zhang J. A Bijection between Lattice-Valued Filters and Lattice-Valued Congruences in Residuated Lattices[J]. Mathematical Problems in Engineering, 2013, 36(8): 4218-4229.

[19] Yang XL, Zhou B, Feng J, Shen PY. Combined energy minimization for image reconstruction from few views[J]. Mathematical Problems in Engineering, 2012, 154630.

[20] Yang XL, Shen PY, Zhou B. Holes detection in anisotropic sensor networks: Topological methods[J]. International Journal of Distributed Sensor Networks, 2012, 8(10): 135054.

[21] H. M. Srivastava, Yunyi Zhang, Lei Wang, Peiyi Shen, and Jing Zhang. A local fractional integral inequality on fractal space analogous to Anderson's inequality[C]//Abstract and Applied Analysis. Hindawi Publishing Corporation, 2014, 46,(8): 5218-5229.

[22] Fan X, Wozniak M, Song H, Li W, Li Y, Shen P. H∞ Control of Network Control System for Singular Plant[J]. Information Technology And Control. 2018, 47(1): 140-50.

[23] Fan, X., Song, H., & Wang, H. Video tamper detection based on multi-scale mutual information[J]. Multimedia Tools & Applications, 2019, 78(19): 27109-27126.

[24] Marcin Wozniak, Robertas Damasevicius, Xiumei Fan, Ye Li, Algorithm Research of Known-plaintext Attack on Double Random Phase Mask Based on WSNs[J], Journal of Internet Technology, 2019, 20(1): 39-48.

[25] Qi Yong. Information potential fields navigation in wireless Ad-Hoc sensor networks[J]. Sensors, 2011, 11(2): 4794-4807.

[26] Liu Shuai, Li Wenjia, Du Dingzhu, Fractal Intelligent Privacy Protection in Online Social Network Using Attribute-Based Encryption Scheme[J], IEEE Transactions on Computational Social Systems, 2018, 5(3): 736-747.