An improved hybrid recommendation algorithm based on feature preference

Jiwen XU¹,a

¹School of Computer Science, Guangdong Neusoft Institute
²Email: xujiwen@nuit.edu.cn

Abstract: In order to solve the problem of data matrix sparsity and cold start in the era of big data. This paper designs a collaborative filtering algorithm, which is an improved hybrid recommendation algorithm based on feature preference analysis. It combines the analysis of user feature preference and item feature, and then uses the traditional collaborative filtering idea to recommend the best scoring object to users. The experimental results show that the accuracy of the proposed algorithm is improved obviously.

1. Introduction

As the Internet industry enters the era of big data, the Internet provides people with massive information resources, which greatly meets the needs of users for various types of information. While bringing great convenience, the problem of network information overload has become increasingly prominent, and the demand for users to find the information they need in the massive information has become more urgent.

In this context, information recommendation technology, as an effective means of information filtering, has gradually become a popular research object in the field of computer science and application, and has been widely used in Internet systems [1]. The recommendation system is aimed at when the user does not have a clear demand scenario, by collecting the user's historical information and the characteristics of the recommended items, analyzing the related characteristics, inferring the user's personalized preference for potential items, and then recommending the best choice to the user. Helping users choose the products they need, while greatly facilitating website users, it also effectively improves the website's operational effects and service quality.

Among many recommendation technologies, collaborative filtering algorithm is currently a relatively mature technology that is widely used [2]. The collaborative filtering algorithm believes that users with similar preferences may generally like the same items. Specifically, it can be divided into three types of collaborative filtering algorithms based on users, item-based and model-based [3]. The advantage of the collaborative filtering algorithm is that it can filter the descriptive content with complex expressions, thereby solving the problem of information filtering that is difficult to perform automatic content analysis, and it also has the function of recommending new information. However, the collaborative filtering algorithm also has its problems and limitations when the amount of data continues to increase. The typical problems are the sparsity of the score matrix and the cold-start problem. In recent years, the academic community has conducted in-depth research and put forward a lot of problems. Effective improvement plan [4].

This article chooses to combine the user's product evaluation content to improve the accuracy of the algorithm. Whether in the information media or e-commerce field, users often post relevant comments. In these comments, there are also implicit affirmative or negative attitudes and attention to products.
These comments are important to users. Websites have important value. Extracting this information can recommend suitable objects for users more accurately.

Aiming at some problems in the current recommendation algorithm, this paper proposes an improved hybrid recommendation algorithm based on the combination of user feature analysis and product feature analysis. After the user enters the system for the first time, first use ICT-CLAS technology to segment the comment text and mark the part of speech, then summarize and quantify the feature words, calculate the item feature weight, and complete the first step of data collection and processing; the next step is to calculate user preference features based on user behavior Value; combined with the traditional collaborative filtering recommendation algorithm, through neighbor user clustering, calculate the user's evaluation scores for other items, and recommend the content of the item with the highest score to the user. After experiments, the results show that this method can improve the accuracy of the recommendation.

2. System outline design

2.1. Overall framework and steps
The improved hybrid recommendation algorithm based on feature preference analysis uses user-product rating data as the initial data source, and divides the entire recommendation process into the following major steps:

1) Preliminary data processing: use Chinese word analysis method to segment a large amount of comment data, and indicate the part of speech of the words at the same time;

2) Feature word induction and quantification: filter the feature words whose frequency reaches a certain threshold, pruning them, and filter out the feature words that do not meet the requirements; quantify the evaluation on the dimension of the feature words, and calculate their weight according to a specific formula;

3) Construct a user preference vector: construct a vector representing the user's feature preference, and calculate the user's similar neighbor set according to the idea of a clustering algorithm;

4) Generate recommended content: Calculate the value of user rating similarity and feature similarity, and push the item with the highest evaluation score to the user.

2.2. Related data
The set of users is expressed as $U = \{U_1, U_2, \ldots, U_n\}$; the set of movie products is expressed as $P = \{P_1, P_2, \ldots, P_n\}$; the product feature set is expressed as $F = \{F_1, F_2, \ldots, F_n\}$; the user characteristic set is expressed as $R = \{R_1, R_2, \ldots, R_n\}$; the matrix of user-product ratings is expressed as matrix $R_{U \times P}$, where $R_{ij}$ represents the user $U_i$, pair of product $P_j$ the evaluation score of $j$; the product-feature association matrix is expressed as the matrix $W_{F \times P}$. Where $w_{ij}$ represents the weight ratio score of the feature keyword $F_i$ to the product $P_j$.

The product-feature score matrix is expressed as matrix $W_{F \times P}$, and the matrix diagram is shown as shown in Figure 1.

| item  | Feature$_1$ | ... | Feature$_j$ | ... | Feature$_m$ |
|-------|-------------|-----|-------------|-----|-------------|
| item$_1$ | $r_{11}$    | ... | $r_{1j}$    | ... | $r_{1m}$    |
| ...  | ...         |     | ...         |     | ...         |
| item$_n$ | $r_{n1}$    | ... | $r_{nj}$    | ... | $r_{nm}$    |

Figure 1 Schematic diagram of product feature matrix
3. New hybrid recommendation algorithm

Feature preference analysis solves the problems of accuracy and limitations of recommendation algorithms by mining the implicit preferences of user comments in the text.

Step 1: Preliminary analysis and processing of review data; the algorithm first uses the Chinese lexical analysis algorithm ICT-CLAS (Institute of Computing Technology, Chinese Lexical Analysis System) to segment movie review texts obtained on the Internet in Chinese [5]. ICT-CLAS is a word analysis system developed by the Institute of Computing Technology, Chinese Academy of Sciences. Its main functions include word segmentation and part-of-speech tagging.

The basic idea of the technology is to first perform atomic segmentation, and then perform N-shortest path segmentation on this basis, find the first N most consistent segmentation results, generate a binary word segmentation table, further perform part-of-speech tagging and complete the word segmentation step. ICT-CLAS system is the current mainstream Chinese lexical analyzer.

Use Chinese word analysis algorithms to mark words as part of speech, such as nouns, verbs, adjectives, and adverbs. Through a large number of observations and statistics and setting thresholds to filter the data, conclusions can be drawn. Feature words are generally nouns and verbs, and emotional words are generally adjectives or adverbs. This completes the collection and analysis of the data of feature words in candidate words and user emotional preference words.

Step 2: Filtering and summarizing feature words; item features generally refer to nouns that describe the nature of objects that are mentioned frequently in reviews [6]. After many attempts at pruning operations, a moderate threshold standard is obtained. The candidate words are pruned based on this standard, and candidates that are not directly related to the performance of the item and those that are less than the specified threshold are filtered, extracted and object features Directly related emotional words to improve the accuracy of the recommendation results.

Since the number of feature words of an item is often large, the dimensionality of the matrix after modeling is too high and the calculation is cumbersome and complicated. Therefore, the filtered feature words can be summarized into several large types to simplify later calculations. The movie feature words extracted from user comments in this paper can be divided into several categories based on similarity.

Step 3: Construct the item feature vector; after summarizing the feature words, the next step is to construct a quantitative item feature vector, and score each feature dimension on a five-point system. The average score on multiple feature dimensions expresses the user’s tendency to evaluate the item. The favorite is 5 points, and the least favorite is 1 point. The feature value of a product on a specific attribute is determined by the joint evaluation of all users on the feature. Thereby, the evaluation scores of the product in several characteristic dimensions are obtained $G(j) = \{f_1, f_2, \ldots, f_n\}$.

Step 4: Calculate the weight of feature words; use TF-IDF technology (Term Frequency-Inverse Document Frequency) to represent the text in the form of a vector based on the frequency of the feature words. When the frequency of keywords in all documents is low, they also appear in some comments The frequency of is higher, and the weight of feature words is higher [7]. Combining web content features and user profile models, the final function can be defined as: where $D$ is the total number of files in the corpus, and $DF(t)$ is the frequency of words appearing in the documents:

$$ u(c,s) = f(Profile,Contents) = tf_{ij} \cdot idf_{ij} = \sum \log \frac{DF(t)}{DF(i)} $$

Using formula (1), the value of u can be obtained, and the larger value can be used as a reference feature with higher weight [8].

Step 5: Construct a user preference vector; the traditional recommendation algorithm only considers the user's overall rating of the item from a single dimension, and the flexibility and accuracy of recommendation are not high. Because different users have different levels of attention to products, this article proposes a vector recommendation method based on user preferences in multiple dimensions, and user preferences will change over time. Through continuous learning, the system can update users in time Eigenvectors. Define matrix R as the user's preference vector. $R = [r_1, r_2, \ldots, r_n]$ user preference categories of a degree of characteristic parameters. The user's feature vector calculation formula is:
where $P$ represents the feature vector of the product. The user's comprehensive rating of the product is the product of the preference vector $R$ and the product category vector $G(f)$ \(^9\).

\[
R(t) = G(f)R
\]  \(^3\)

Step 6: Calculate user similarity through clustering; the clustering algorithm divides the object set into different subsets. Objects in the same subset have higher similarity. This article uses this idea to put users with high similarity in a subset, in order to reduce the amount of calculation of the user's neighbor set \(^10\). Randomly select $k$ objects from the user set as the center point of the initial subset; calculate the distance between the user feature vector and these subsets, and re-divide the feature vector according to the principle of minimum distance; recalculate the mean value of each subset as the center point of the subset; Repeat the first two parts to complete the user feature vector clustering.

User similarity can be divided into two categories-user preference similarity and user rating similarity \(^11\). User preference similarity can be calculated using the angle cosine value. The cosine similarity calculation user preference vector similarity formula is:

\[
sim(r_i, r_j) = \cos(r_i, r_j) = \frac{r_i \cdot r_j}{\|r_i\|\|r_j\|}
\]  \(^4\)

among them, $r_i$ and $r_j$ represents the user's preference vector, which $\sim(r_i, r_j)$ represents the degree of similarity in the preferences between the two. The larger the calculation result, the closer the two users' preferences are \(^12\).

Since users with similar preferences have similar scoring behaviors, the user's rating of the same item is used to calculate user similarity, and the most commonly used Pearson Correlation Coefficient formula is selected. The formula is as follows:

\[
sim(i, j) = \frac{\sum(s_{ic} - S_c)(s_{jc} - S_j)}{\sqrt{\sum(s_{ic} - S_c)^2} \sqrt{\sum(s_{jc} - S_j)^2}}
\]  \(^5\)

among them, $s_{ic}$ and $s_{jc}$ represent the ratings of user $i$ and user $j$ on the same rating item $c$, and $S_c$, $S_j$ represents the average rating of the user on the item \(^13\).

Step 7: Comprehensive score calculation and item recommendation; use the user-based collaborative filtering algorithm to calculate the target item's score for a specific product through a neighbor set similar to the target user, select neighbor users with greater similarity to form a neighbor set, and use a weighted average. The strategy, according to formula (7), calculates the recommendation result by weighting the scores of the neighbors on the items

\[
P(u, i) = S_i + \frac{\sum_{m \in \text{neighbors}}(s_{mi} - S_m) \cdot \text{sim}(n,m)}{\sum_{m \in \text{neighbors}} \text{sim}(n,m)}
\]  \(^6\)

where $S_i$ represents the average score of the item, $s_{mi}$ represent the ratings of the neighbor users for the item $i$, $S_m$ represents the average score of all the neighbor users, and $\text{sim}(n,m)$ represents the comprehensive similarity between the target user $u$ and the neighbor user $m$ \(^14\). Recommend the top $k$ objects with the highest recommendation prediction score to the user.

4. Experimental analysis

The average difference formula is currently the mainstream evaluation standard for evaluating the performance of a recommender system. The actual score and the predicted score are calculated as the absolute average error MAE (Mean Absolute Error) to compare the degree of deviation between the predicted score and the actual score. In order to judge the accuracy of the recommendation, the smaller the average difference, the better the effect of the recommendation system \(^15\). The algorithm predicts
the score set \( \{s_1, s_2, \cdots, s_n\} \), the user’s actual score set is \( \{q_1, q_2, \cdots, q_n\} \), and the formula for the average difference is as follows:

\[
MAE = \frac{1}{n} \sum_{i=1}^{n} |s_i - q_i|
\]

A web crawler is used to crawl the user's evaluation information and scoring data of 1,299 products in an Internet movie database, and the data is randomly divided into two parts, of which 80% is classified as the training set and 20% is classified as the test set. The experimental results show that the traditional user-based or content-based collaborative filtering algorithm has a larger absolute average error MAE value and lower recommendation accuracy. In comparison, the new recommendation algorithm has a smaller MAE value, and the recommendation accuracy is significantly improved. Experimental data such as shown in Figure 2.

![Figure 2](image)

**Figure 2** Comparison of algorithm recommendation accuracy

The advantage of the new hybrid recommendation algorithm based on feature preference analysis is that it first analyzes the review text and extracts the key feature information. The core idea is to comprehensively utilize the feature analysis methods of both users and products, and then through the user-based collaborative filtering algorithm. The final data is processed, so the new algorithm has obvious advantages compared with the traditional recommendation algorithm, especially when the number of user goals is low, the accuracy improvement effect is more obvious.

5. **Conclusion**

This paper proposes an improved hybrid recommendation algorithm based on feature preference analysis, which calculates item feature vectors and user feature preference vectors, and recommends the items with the highest ratings to users through the clustering of neighbor users similar to the target user and the use of collaborative filtering recommendation algorithms. Which alleviates the matrix sparsity problem and cold start problem in traditional recommendation algorithms. The experimental results show that the improved hybrid recommendation algorithm based on feature preference analysis effectively improves the accuracy of the recommendation results.

**References**

[1] Herlocker, JL, Konstan, JA, Borchers, A., et al. An Algorithmic Framework for Performing Collaborative Filtering. Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Berkeley, 15-19 August 1999, 230-237.

[2] Xu Hailing, Wu Xiao, Li Xuedong, et al. Comparative Research on Internet Recommendation Systems[J]. Journal of Software, 2009, 20(2): 350-362.

[3] Sun Wenshuang. Principal component method of ridge regression parameter selection[J]. Journal of Yunnan University (Natural Science Edition), 1991, 4: 1.
[4] Bobadilla, JS, Ortega, F., Hernando, A., et al. (2012) A collaborative Filtering Approach to Mitigate the New User Cold Start Problem. Knowledge-Based Systems, 26, 225-238.
[5] Xiang Liang. Recommendation system practice[M]. Beijing: People's Posts and Telecommunications Press, 2012.
[6] Yang Yang, Xiang Yang, Xiong Lei. Collaborative filtering recommendation algorithm based on matrix factorization and user nearest neighbor model[J]. Journal of Computer Applications, 2012, 32(2): 395-398.
[7] Wu, HC, Luk, RWP, Wong, KF, et al. (2008) Interpreting tf-idf Term Weights as Making Relevance Decisions. ACM Transactions on Information Systems, 26, 13.
[8] Zhao Dan, Cai Shu. Product recommendation model based on online shopping reviews[J]. Journal of Southwest University for Nationalities: Natural Science Edition, 2012, 38(3): 471-474.
[9] Deng Ailin, Zuo Ziyue, Zhu Yangyong. Collaborative filtering recommendation algorithm based on item clustering[J]. Small Microcomputer System, 2004, 25(9): 1665-1670.
[10] Lu Zhubing. Research on collaborative filtering recommendation strategy based on trust relationship [D]: [Master's thesis]. Chongqing: Southwest University, 2008.
[11] Liu Jia. Score prediction based on matrix factorization collaborative filtering algorithm[J]. Science & Technology Innovation and Application, 2016(5): 86-86.
[12] Zhang Shijie, Yao Jianping. A new collaborative filtering recommendation model based on comment mining[J]. Technological Innovation and Productivity, 2013(3): 52-54.
[13] Chen, J. and Yin, J. (2007) A Collaborative Filtering Recommendation Algorithm Based on Influence Sets. Journal of Software, 18, 1685-1694.
[14] Zhimin Chen, Zhiqiang Li. Collaborative filtering recommendation algorithm based on user characteristics and item attributes[J]. Journal of Computer Applications, 2011, 31(7): 1748-1750.
[15] Liu Yu, Zhang Guangyi, et al. Recommendation algorithm based on tag combination effect in collaborative tag system[J]. Journal of South China University of Technology: Natural Science Edition, 2013, 41(9): 65-70.