Research on Personalized Recommendation System Based on Association Rules

Song HaiYan¹, Zhang Huan², Xing ZhiHao
¹Hu Zhou Vocational & Technical College, 299 Xuefu Road, Huzhou City, China
²North China Electric Power University, No.619, Yonghua North Street, Baoding City, China
paperbo@sina.com

Abstract. With the rapid development of network technology, the amount of information is rapidly expanding, and the amount of various data has become huge and scattered. Using traditional keywords to retrieve search data has become quite time-consuming and difficult to focus on. Traditional search engines have been unable to help people effectively solve this problem, so personalized recommendation systems have emerged. In view of the user's behavior and preferences, this article makes a detailed comparison of different recommendation technologies. It focuses on analyzing the applicability of the current popular content-based, association rules, and collaborative filtering recommendation technologies to the current needs. The technical scheme and system structure of recommendation system based on association rules are proposed, and aiming at the shortcomings of traditional recommendation system algorithms (the algorithm's sparsity problem) and the current cold start problem of the recommendation system. Based on the MovieLens data set, a series of evaluation indicators are set up, comparative experiments are carried out, and the effectiveness of the proposed improved algorithm is analyzed.

1. Introduction

With the development of electronic information technology, network technology, and Internet technology, human life style has undergone revolutionary changes. Electronic information technology makes it possible to store and process massive amounts of information. Network technology also enables data processing results to be distributed to different systems, which greatly increases the amount of information stored, the degree of sharing, and the speed of information processing. The development of the Internet has expanded this trend from a small area to the entire planet. At present, these technologies have now penetrated into all areas of people's study, work and life. People can get rid of the constraints of time and space, and obtain various services conveniently and efficiently. However, in the face of the vast network information space, users need to spend a lot of time searching for information, which also greatly reduces the utilization rate of information, leading to the phenomenon of "information overload". Personalized recommendation technology came into being. The essence of recommendation is to replace users in evaluating resource objects and help users quickly find their favorite products. Although the personalized recommendation system has been applied in the field of e-commerce, there are still many problems that need to be solved urgently, such as cold start, sparsity. These problems have attracted widespread attention from scholars.
2. Technology of Personalized Recommendation System

With the rapid development of network technology and explosive growth of information, the problem of information overload is becoming more and more serious. It is difficult for users to quickly find the information they need. Some important information is submerged in the sea of information and becomes isolated information. The amount of information is rapidly expanding, and various types of data have become huge and scattered, and users using traditional methods to search for data have become time-consuming and difficult to focus on. In the case that traditional keyword retrieval cannot help people solve this problem, a recommendation system was created.

2.1. Recommended Method Classification

2.1.1. Recommendations Based On Content. By analyzing the historical behavior of users in the system, that is, describing and scoring existing projects. On this basis, according to the investigation and analysis of the items processed by the user, the feature model of the item and the personal characteristics of the user can be established. The key step of the content recommendation process is to match the characteristics of the user's personal information with the characteristics of the content object, and the recommendation result reflects the user's preference for a certain object. If personal data can accurately reflect user preferences, content-based recommendation. First, Item Representation. Second, Profile Learning. Third, Recommendation Generation. Figure 1 is a high-level structure of a content-based recommendation system. Each of the three stages of the recommendation process is completed by an independent component.

2.1.2. Recommendations Based On Collaborative Filtering. The collaborative filtering recommendation system is based on the selection of other users to recommend a document to a user, regardless of the content of the document. It is not ideal if a system fails to respond when the user provides fuzzy query conditions. Therefore, some systems integrate and summarize user preference...
information based on statistics to provide personalized recommendations. The system will analyze the preferences and interests of different users, as a basis to classify users into different types of communities. Then, these systems provide users in the same community with objects that match their preferences. By comparing the preference information of the new user with the existing community preference information, and uploading this user's preference to the community's preference database, the system can summarize the user's preference information.

2.1.3. Mixed Recommendation. No matter what kind of recommendation algorithm, there are advantages and disadvantages. In the process of practical application, in order to improve the performance of the recommendation system, the inferior part of an algorithm can be replaced by a strong algorithm to form a mixed recommendation. For example, based on content recommendation and collaborative filtering recommendation, the content recommendation algorithm is added to the collaborative filtering algorithm or the collaborative filtering algorithm is added to the content recommendation algorithm. Due to the diversity of the algorithm itself, there is no situation where any two algorithms can be combined at will. After continuous exploration, there are seven main types of mixed recommendations, Weight, Switch, Mixed, Cascade, Feature Combination, Feature Augmentation, Meta-Level.

2.2. Recommended Technology Options
The basic idea of content-based recommendation is that a user will prefer the products they have bought and similar products. So the core of the algorithm is to analyze the attributes of each product, and calculate the products that should be recommended on this basis. To analyze the similarity of products, you first need to describe the content of the product. The description is generally through the definition of a set of product attributes, and similarity calculations through different attributes of the product. I learned from the literature that there are two main methods of calculation, and vector approximation calculation and conditional similarity. Likelihood can intuitively show how well the algorithms supporting different recommendation methods cover user preferences. Collaborative filtering and association rules do not have the above technical limitations, and are the most successful and widely used recommendation technology. The biggest feature of these two algorithms is that they do not need special data source support and can make full use of existing resources. Both algorithms can provide users with highly accurate recommendations through historical records. These recommendations can not only satisfy personalized preferences, but also predict user behavior to a certain extent. The recommendation system introduced in this article will use a recommendation algorithm combined with association rules.

3. Recommendation System Based on Association Rules
3.1. Description of The Basic Concepts of Association Rules
Association Rules were first proposed by Agrawal (1993) and others. They mainly analyze the problem of Basket Analysis, mainly to discover the association rules of different commodities in the transaction database. Association Rules reflect the relationship between a specific data, used to reveal the unknown interdependence between data and data. Its task is to find a large number of interesting and relevant connections between data and items in a support-confidence framework based on a transaction database, and generate all association rules with confidence and confidence higher than the minimum support and minimum confidence given by the user, respectively. Two problems with its algorithm design, first find all Item Sets with support greater than or equal to the min_sup support, these item sets are called Frequent Item Sets. Second, use the frequent itemsets found in step1 to generate expected rules. Three metrics for association rule mining, Support, Confidence, Lift.

3.2. Principle and Process of Association Rules
The evaluation criteria of association rules are mainly support and confidence. The two thresholds of
support and confidence are two important concepts for describing association rules. For convenience, the minimum support threshold is denoted as $\text{min}_\text{sup}$, and the minimum confidence threshold is denoted as $\text{min}_\text{conf}$. The minimum support indicates the lowest importance of the item set in a statistical sense. The minimum confidence indicates the minimum reliability of the rule. Suppose it is a collection of "n" different data items. Given a data transaction set "T". Each transaction record "t" is a non-empty subset of "I". That is, every transaction record corresponds to a unique identifier TID. For any non-empty itemset, if record "t" contains "X", it is said that the record "t" supports the item set "X". For the entire data set T, the support of X is defined as the proportion of records containing X in the data set T. The formula is as follows.

$$D\sup p(X) = \frac{\|X\|}{|T|}$$

Where $\|X\|$ is the number of records of X contained in the data set T, $|T|$ represents the number of all records in T. Obviously, if $|X|=k$ (that is, $|X|$ represents the number of data items in X), then X is called a k-item set. If the support of X is greater than the given minimum support threshold, then X is called the frequent set. The expression of the association rule is as follows, where X and Y are non-empty itemsets, and X and Y do not intersect. The support of association rules is defined as follows.

$$D\sup p(X \implies Y) = D\sup p(X \cup Y) = \frac{\|X \cup Y\|}{|T|}$$

Where X is the former term, and Y is the latter term. The confidence level is the former term is used as the denominator.

$$D\text{conf}(X \implies Y) = \frac{\|X \cup Y\|}{\|X\|} = \frac{\|X \cup Y\|}{\|X\|} = \frac{\|X \cup Y\|}{\|X\|}$$

Where $\|X \cup Y\|/|T|$ and $\|X\|/T$ represent the number of union and X records in the data set T, respectively.

If the required association rule support is greater than the minimum support, and the confidence is greater than the minimum confidence, it is called a qualified association rule.

It should be noted that for an association rule, the work of calculating support is mainly counting operations, requiring traversal of the scanned data set T.

### 3.3. Association Rule Algorithm

Association rule mining algorithm is the main content of association rule mining research. So far, many efficient association rule mining algorithms have been proposed. The most famous association rule discovery method is the Apriori algorithm proposed by R. Agrawal. The Apriori algorithm mainly consists of two steps. The first step is to find all the data item sets in the transaction database that are greater than or equal to the minimum support specified by the user. The second step is to use frequent itemsets to generate the required association rules. According to the minimum confidence level set by the user, the strong association rule is finally obtained. Identifying or discovering all frequent item sets is the core of the association rule discovery algorithm. There are two key steps in the core idea of the Apriori algorithm, the connection step and the pruning step. The meaning of the connection step is to find out $L_k$ (frequent k-item set), connect with itself through $L_{k-1}$, and generate a candidate k-item set, which is denoted as $C_k$; the elements of $L_{k-1}$ are connectable. The meaning of the pruning step is that $C_k$ is a superset of $L_k$, that is, its members may or may not be frequent, but all frequent itemsets are included in $C_k$. We scan the database to determine the count of each candidate in $C_k$, thereby determining $L_k$(all candidates whose count value is not less than the minimum support count are
frequent and thus belong to Lk). However, Ck may be very large, so the amount of calculation designed is very large. In order to compress Ck, the Apriori algorithm is used: any infrequent (k-1) itemsets cannot be a subset of frequent k itemsets. Therefore, if the (k-1) item set of a candidate k item set is not in Lk, the candidate item may not be frequent, so it can be deleted from Ck. This subset test can be quickly completed using the hash tree of all frequent itemsets.

3.4. Recommendation Mechanism Based On Association Rules
The framework of the recommendation mechanism is shown in Figure 2. The knowledge in the preference group is preprocessed to form the basis of the entire recommendation mechanism. In the application process of data mining technology, association rules and grouping knowledge are produced. When a user service request is sent, the system will identify the user's preference and find a rule that suits his preference to make predictions after identifying the request. When the rules are activated and the appropriate knowledge is found, the system will provide the results to the user, and the user's personalized needs can be met.

![Figure 2 Recommendation Mechanism Framework](image)

For users, the efficiency of online recommendation is an important factor. In order to achieve the goal of high efficiency, neural network technology is applied to user type classification. Once the recommendation system has been learned, it can classify any online user in real time through the user's personalized attributes. Another key factor is to summarize the knowledge in different user groups. The association rules of data mining are often used to find the relationship between the items selected by
the user and the user behavior. By integrating these two technologies, it will be easy to get user behavior knowledge.

User preferences will be analyzed and preprocessed to generate ethnic group information. This information combined with historical transaction data can generate the characteristics of different user groups and their association rules. When the user's service request is sent, the system will identify his preference type, and extract the matching association rules from the knowledge base to recommend the user. Based on such a method and process, the system can give timely online recommendations to different users.

The goal of the automated recommendation system is to recommend relevant information to users based on user types and items selected by users. The first step of this mechanism is to preprocess user information, clarify user attributes and user historical data. Information cleaning, merging, conversion, etc., The experimenter organizes all the information into a sub-data set suitable for analysis, and then analyzes it with a suitable algorithm.

4. Simulation Experiment Performance Analysis

4.1. Evaluation Indicators and Data Sets

So far, there is no algorithm that can be applied to all recommendation systems. How to measure the quality of recommendation algorithms and whether a sound evaluation system can be established is particularly important. The current method used by mainstream recommendation systems is to select recommendation accuracy as a reference to measure the pros and cons of the algorithm. The more widely used ones are prediction accuracy and classification accuracy. Other evaluation indicators can be used as auxiliary indicators in order to make the evaluation of the recommendation system more comprehensive. The following introduces these more important indicators in the recommendation system.

4.1.1. Forecast Accuracy.

If the user's historical ratings are recorded in the recommendation system, then the prediction accuracy can be used as an evaluation method for this recommendation system. The most classic evaluation method of prediction accuracy is the Mean Absolute Error (MAE). The main idea is to compare the average absolute error between the forecast score and the user's actual score. The error value is inversely related to the recommended accuracy. The smaller the error, the higher the accuracy. The formula is as follows.

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |\hat{v}_{iu} - v_{ui}|$$

In the above formula, "n"represents the total number of users "i" in the recommendation system, $v_{ui}$ represents the user's true rating, and $\hat{v}_{iu}$ represents the system's predicted score. The average value of the accuracy of all users of the system is the accuracy of the recommendation algorithm.

4.1.2 Classification Accuracy.

The difference between classification accuracy and prediction accuracy is that it is only used to determine the proportion of users who are satisfied with the objects recommended by the system, regardless of whether the algorithm score is correct. Commonly used classification accuracy indicators mainly include accuracy rate, recall rate and hit rate. The following also briefly introduces these indicators.

(1) The accuracy rate represents the percentage of the ratio of the number of objects that meet the user's preference to the number of all recommended objects in the recommended list given. The accuracy rate reflects the target user's preference for the recommended resource object. The formula is as follows.
The recall rate is a measure of coverage, which measures the percentage of the number of objects that meet the user's preferences in the recommended list given to the number of objects that they like. The recall rate reflects the probability of the target user's preference in the recommendation system. The formula is as follows.

\[ P = \frac{N_n}{N_s} \]

(2) The hit rate represents the percentage of the ratio of the number of hits to the length of the list given the length of the recommended list. The formula is as follows.

\[ R = \frac{N_n}{N_n + N_m} = \frac{N_n}{N_r} \]

In the test, if the user \( U_i \) that appears chooses object \( O_j \), and object \( O_j \) also appears in the recommendation of user \( U_i \). In the list, it shows that the recommendation algorithm is forbidden (User \( U_i \)-Object \( O_j \)). The fate rate increases as the length \( L \) of the recommendation list increases, and the maximum value is 1.

4.2. Data Centralized Experiment and Analysis

4.2.1. Data Preprocessing. In order to eliminate the influence of abnormal data, preprocessing of the data set is a prerequisite. Remove objects less than or equal to once in the data set. Ensure that each user selects at least one object, and each object has been selected by at least two users. After preprocessing, the data set is shown in Table 1.

| Types                   | MovieLens Dataset |
|-------------------------|-------------------|
| Amount of users         | 95                |
| Number of Objects       | 334               |
| Number of user-object pairs | 1643             |

4.2.2. Forecast Accuracy Analysis. Evaluate the effectiveness of the proposed algorithm for improving the recommendation results by comparing the prediction accuracy of the association rule-based recommendation and the traditional collaborative recommendation algorithm proposed above.

The prediction accuracy of the two algorithms compared here under different numbers of users. The horizontal axis in Figure 3 represents the number of users, each time increasing by 5 units, from 5 units to 95 units in turn; the vertical axis represents the MAE of different algorithms value. The analysis steps are as follows.

(1) Randomly selected as a movie that needs to be scored as a subject.
(2) After the selected experimental subject has been rated by a user, the user's score vector is calculated using the scores of other subjects given by the user.
(3) Use the user score vector to form the attribute score matrix, and then use the traditional recommendation algorithm to calculate the scores of the experimental subjects under different numbers of users, and then use the algorithm based on association rules to repeat the above steps.
(4) After completing step 3, there will be a difference between the predicted value and the actual value. Use this difference to calculate the predicted average absolute error MAE under different numbers of users. The calculation result is shown in Figure 3.

![Figure 3 MAE Change Diagram of Two Algorithms](image)

It can be seen from the figure that when the number of users increases from 5 units, the MAE values of the two recommendation algorithms show a rapid downward trend. As the number of users further increases, the MAE value remains stable after it drops to a certain level and starts to fluctuate. This shows that the improvement of the prediction accuracy of the recommendation algorithm is limited, and the calculated predicted value is almost only close to the actual value under normal circumstances. This is because the user will be affected by the mood and scene at the time, and will not reflect 100% of his true preference for the product. From the analysis of the change in the slope of the broken line in the figure, it can be found that although the MAE value based on the association rule has small fluctuations. But in most cases, its MAE value is lower than the prediction result of traditional recommendation algorithm. The number of users reached 25 units, and the association rule recommendation basically reached a stable state. When the number of users reaches 30 units, the traditional recommendation algorithm stabilizes. This shows that association rule recommendation can get the same or even better results of traditional algorithms through less user rating data and smaller calculations. This has a positive effect on the cold start problem of the recommendation system caused by the recommendation algorithm.

4.2.3. Classification Accuracy Analysis. In the experiment, we compared the performance of the two algorithms in prediction accuracy. Let's compare their performance in classification accuracy. Since classification accuracy has many evaluation indicators, a more intuitive hit rate is selected for analysis here. Figure 4 is a simulation flow chart of hit rate.
In the MovieLens data set, as shown in Figure 5, the horizontal axis is the recommended list length \( L \), which grows from 10 units to 100 units, and the interval is 10 units. The vertical axis represents the recommended accuracy (that is, the hit rate).

Figure 5 Hit rate under different recommended list lengths in

It can be seen from Figure 5 that the algorithm based on association rules has better effect, and is different from the traditional recommendation algorithm, which increases with the increase of the recommendation list length \( L \), and then gradually stabilizes, and its recommendation accuracy reaches the highest value earlier.

As the recommended list length \( L \) increases, it decreases and tends to stabilize. Analyzing the reason for this situation, it may be that the coverage rate of the algorithm proposed in the article is small. After the length of the list reaches a certain level, the recommended items have been or nearly all recommended. Increasing the length of the recommendation list will naturally reduce the recommendation accuracy. But at the same time, it should be noted that after the algorithm based on association rules is stable, the recommendation accuracy is still better than traditional recommendation algorithms.

Introduce the idea of association rules in the recommendation system, use the correlation between items to generate an association matrix, use the association matrix to generate a candidate item set for users, and separate the prediction score from the candidate item set. Make the candidate item set more in line with the user's interest orientation, use less data to effectively recommend the user, and play a positive role in the cold start problem of the recommendation system. The quality of the project candidate set has been improved, the size of the candidate set has been reduced, and the speed of predicting scoring and generating recommendation lists has been improved. This shows that data
sparsity issues and system scalability issues can be mitigated to a certain extent by using recommendation based on association rules.

5. Research Conclusion Analysis
The rapid development of information technology has made the recommendation system more and more significant, and Internet users are increasingly hoping to obtain valuable information for themselves in a short period of time. This research combines collaborative filtering and association rule technology to design a new user-oriented recommendation mechanism. That is, an association rule algorithm is used to generate the association set of all products, and the strong association rules between the products are mined as the recommendation rule. Finally, according to the information input by the user, appropriate related products or information are recommended to the user. The advantages of this recommendation mechanism are as follows.

(1) Community characteristics. It can analyze the different personal backgrounds of users, perform clustering, and then generate different communities, and then recommend product information suitable for each community.

(2) Associated features. Combining association rule algorithms to discover the relevance of products for product recommendation, its depth and breadth are more referential than traditional recommendation methods.

(3) High recommendation efficiency. When the algorithm digs out the association rules, the system can immediately recommend the appropriate information to the user, so it is very suitable for dynamic changes and rapid response environment.

References
[1] Xu Hai-yan, Wu Xiao, Li Xiao-dong, Yan Bao-ping. Comparison Study of Internet Recommendation System[J]. Journal of Software, 2009.20(2):350-262
[2] Adomavicius G, Tuzhilin A. Expert-driven validation of Rule-Based User Models in personalization application [J]. Data Mining and Knowledge Discovery, 2001, 5(1 —2) : 33—58
[3] Adomavicius G, Tuzhilin A. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions [J]. IEEE Trans on Knowledge and Data Engineering, 2(X)5, 17(6) : 734—749
[4] O’Donovan J, Smyth B. Trust in recommender systems [C]. in Proceedings of the 10th International Conference on Intelligent User Interfaces. New York: ACM Press, 2005,167-174
[5] Wang Wei. Association Rule Mining and Change Mining Based on the Concept Lattice[D]. Shan Dong University, 2012
[6] Luo Hong-ying. Improved Method of Correlation Mining Based on Mathematical Model of Calculus Classification[J/OL]. Modern Electronics Technique, 2019(08):135-139[2019-05-04]. https://doi.org/10.16652/j. issn. 1004-373x.2019.08.030.