Automatic Clustering Algorithm for Movie Recommendation Based on LFM Model

Kuihe Yang, Shan Liu, Tao Liu and Xiang Wang

ABSTRACT

In order to enhance audience satisfaction and recommend them different types of movies, it is necessary to point out the concept, latent factor model (hereinafter referred to as LFM). By using clustering analysis technology, users can be divided into different clusters according to the similarity principle, and then according to the evaluation information of the users in the same cluster. LFM algorithm based on the user's behavior to generate implicit feedback, automatic clustering of the user's interest in the class and the hidden category. Because the clustering process can be carried out offline, the clustering process does not affect the response speed of the recommendation system. The figures show that audience will classify the movies automatically. Experiments have showed that the method is effective and accuracy.

INTRODUCTION

In the internet environment, the over populate information is challenging e-commerce. The users have to spend more and more money in getting the information which they really need [1]. And many of them cannot afford to the cost. The recommender systems, as an information filtering technology, has gradually become an effective tool to solve the problems of information over load on the internet [2]. The exiting recommendation technologies mainly base on content, collaborative filtering, knowledge and mixed reasons. Some recommendation technologies have been widely used in large e-commerce systems, such as Amazon and eBay.
Collaborative filtering recommendation algorithm[3, 4] is the most widely used and the most successful personalized recommendation technology in e-commerce recommendation system. Collaborative filtering algorithm can be subdivided into two types based on neighborhood algorithm and model based algorithm[5]. The neighbor based collaborative filtering algorithm can be divided into user based collaborative filtering and collaborative filtering based on collaborative filtering.

In addition, research has indicated that the project based algorithm is generally superior to the user based algorithms in performance[6]. Based on the model of collaborative filtering recommendation is to use machine learning and data mining algorithms, such as training data is used to study identification of complex mode, learning model is obtained, and then based on the learning model in intelligent prediction data sets[7]. However, in the traditional method based on collaborative filtering, the different size of the project cannot be taken care.

In this paper, to latent semantic model as the foundation, the implicit semantic model, introduce personalized movie recommendations platform LFM automatic clustering algorithm, which can be based on user behavior item automatically clustering, which is the item into different categories / topics that / category possible understood as the user's interest.

RELATED WORK

Prediction is one of the important tasks in the recommendation analysis. In general, the commonly used method is based on the user's registration information and historical behavior. Jae Lee Kyung and Chang Woojin using Bayesian network to study the influence factors of the box office, established the related prediction model. Marshall P is presented using the release of the movie has accumulated audience number sum, in the use of multiple linear regression algorithm to predict first week number of viewers, use the first week of the number of viewers to predict a few weeks after the audience the numbers. Independent variable in the model including the number of search queries and film clicks.

In this paper, to latent semantic model as the foundation, the implicit semantic model, introduce personalized movie recommendations platform LFM automatic clustering algorithm, which can be based on user behavior item automatically clustering, which is the item into different categories / topics that / category possible understood as the user's interest.

RECOMMENDATION ALGORITHM BASED ON LFM CLASSIFICATION MODEL

latent semantic CF models(LFM) is an algorithm published by Funk Netflix in Prize Simon. It is a kind of latent semantic analysis technology. In recent years, it is always applied to the project of automatic clustering of some recommendation systems.
LFM Automatic Clustering Algorithm

Based on clustering analysis recommendation, the users can be divided into different clusters according to the similarity principle, and then according to the evaluation information of the users in the same cluster.

In the visible user object comes down to 3 categories, does not mean that the user only like the 3 categories, the object of the other categories of interest is not. That is, the user needs to understand the user's interest in all categories. For a given class, it is required to determine the weight of each object in this class. Weights help determine which objects are recommended to the user. Using LFM to the modeling can be obtained after such as the model shown in Figure 1: (assuming the data from three user, four item, the classification of the LFM modeling the number of 4).

R is the user-item matrix, the matrix representation is the interest value $R_{ij}$ of user i of item j, this is the desired value. Here there is no need to worry about the angle classification and classification of particle size, k class represents the number, LFM algorithm from data extract several themes, as the bridge between the user and item, the R matrix representation for P matrix and Q matrix multiplication. The matrix P is user-class matrix, matrix valued $P_{U,k}$ represents the degree of user i interest of class j; Q matrix class-item matrix of the matrix values, $q_{ij}$ said item j in class i weight, the higher the energy as a representative of the class. So LFM according to Figure 1 to calculate the user's score calculation of each project.

$$R_{ui} = P_u Q_i = \sum_{k=1}^{K} P_{U,k} Q_{k,i}$$

Among them, $P_u$ and $Q_i$ are the characteristic vector of model parameters, $P_{U,k}$ said users interested in u and the k implied the relationship between the characteristics of $Q_{k,i}$, the relationship between the k features and hidden project i. The parameters $P_u$ and $Q_i$ can use the loss function to get training:

$$C = \sum_{(u,i) \in K} (R_{ui} - \sum_{k=1}^{K} P_{U,k} Q_{k,i})^2 + \lambda\|P_u\|^2 + \lambda\|Q_i\|^2$$

In the formula (2) $\lambda\|P_u\|^2 + \lambda\|Q_i\|^2$ is used to prevent over fitting of the regularization term, according to the specific application scenarios need lambda repeated experiments. The formula (2) is used to calculate the partial derivative of $P_u$ and $Q_i$ by the steepest descent method:
After that, the iterative updating and using stochastic gradient descent method are used:

\[
\frac{\partial C}{\partial P_{uk}} = -2(R_{ui} - \sum_{k=1}^{K} P_{uk}Q_{k,i})Q_{k,l} + 2\lambda P_{uk}
\]

(3)

\[
\frac{\partial C}{\partial Q_{kl}} = -2(R_{ui} - \sum_{k=1}^{K} P_{uk}Q_{k,l})P_{uk} + 2\lambda Q_{kl}
\]

(4)

After that, the iterative updating and using stochastic gradient descent method are used:

\[
P_{uk} = P_{uk} + \alpha((R_{ui} - \sum_{k=1}^{K} P_{uk}Q_{k,i})Q_{k,l} - \lambda P_{uk})
\]

(5)

\[
Q_{k,l} = Q_{k,l} + \alpha((R_{ui} - \sum_{k=1}^{K} P_{uk}Q_{k,l})P_{uk} - \lambda Q_{k,l})
\]

(6)

Among them, it is the learning rate, the bigger the alpha, the faster the iterative descent. In the process of iteration, with the constant approximation of the result and the ideal value, the. Like alpha and alpha, both need to be repeatedly tested according to the actual application scenarios. Continuous optimization of the final results by formula (5) and formula (6) iterative calculations until it exceeds the number of iterations (number of iterations that ancestors setting) or threshold condition is satisfied.

**Latent Semantic Movie Recommendation Algorithm Based on Description**

To sum up, to implement LFM, according to the data set to initialize the P and Q matrix. To determine the 4 parameters: the classification number F, the number of iterations of N, learning rate alpha, regularization parameter. Latent semantic movie recommendation algorithm is described as algorithm 1 shows based on.

**Algorithm 1:** The movie recommendation algorithm based on Latent Semantic

**Input:** \( R_{ui}, K, \lambda, \alpha, N, E \);

**Output:** \( P_{u,k}, Q_{k,l} \);

**Step:**
def LFM(user_items, F, N, alpha, lambda):
\[
P_{u,i}, Q_{i,j} = \text{Init Model}(R_{UJ}, K)
\]
2. for step in range(0, N)
    for user, items in \( R_{UJ} \)
        err = err - Predict(user, item)
        for f in range(0, K)
            P[user][f] += \alpha \ast (eui \ast Q[f][item] - \lambda \ast P[user][f])
            Q[f][item] += \alpha \ast (eui \ast P[user][f] - \lambda \ast Q[f][item])
        End if err<E
    End for
End for

**EXPERIMENT**

LFM is used to classify the film, while the use of expert classification and LFM model to carry out a comparative test.

**Data**

With the rapid development of social media, more and more people in the vertical media to share their views, such as the broad bean and IMDB. In this paper, the box office and movie details (type, producer, etc.) are provided by the Chinese film box office network. In detail, including the 2015 box office movie data set hundred in 100 films released in China. In addition, it consists of 36 foreign films and 64 Chinese film composition. Moreover, the 36 foreign films including 745996 comments and 292142 users, and in the 64 China film contains 1365680 comments and 583626 users.

**Simulation Experiment**

According to the Chinese film box office market, the total data set in this experiment is divided into two parts: foreign and domestic.

For the data set of 100 films and the number of users to renumber, while the system will automatically assign a unique user ID (item ID) number. When the user logs in, the system automatically records the interaction between the user and the object. The system records the user number of the user, and the interactive movie number. At the same time, according to the release of the distributor of the film genre film label information as TABLE 1.

Though the producers have classified movies, different people have different standards. Let’s take Monster Hunt (Zhuo Yao Ji) as an example. According to the Chinese film box database, it is labeled with plot, comedy and fantasy. While on douban.com, it is labeled with comedy, fantasy and costume. The different standards
cause the difference. Upon the survey, we record the movies which are popular among the audience in the following form.

Through Table 2, I can know that ID236526 likes comedies most. If we classify them more detailed, they belong to the black humor. We also can classify the audience by the movies they watch. And then we can recommend the same kind movies.

In order to prove the validity, I use LFM automatic clustering algorithm to figure out some movies which rank well. For different dataset, the unknown parameters in Table 2 and Table 3 are different. Many experiments are needed for getting more unknown parameters through which we can obtain the best function and effective learning speed. The unknown parameters in Table 3 are constantly changing. The classification results of LFM model are shown in Table 3. And LFM model classify them. Although the group is just a collection of films and has no specific meaning. Each group has some potential meanings.

<table>
<thead>
<tr>
<th>Table 1. Film Label Information.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Film</strong></td>
</tr>
<tr>
<td>Monster Hunt</td>
</tr>
<tr>
<td>Furious 7</td>
</tr>
<tr>
<td>The Ghouls</td>
</tr>
<tr>
<td>Gang Jiong</td>
</tr>
<tr>
<td>Avengers: Age of Ultron</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 2. User Interest Annotation Information.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ID</strong></td>
</tr>
<tr>
<td>236526</td>
</tr>
<tr>
<td>365968</td>
</tr>
<tr>
<td>386269</td>
</tr>
<tr>
<td>533687</td>
</tr>
<tr>
<td>699823</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 3. LFM Model Foreign Film Classification.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1</strong></td>
</tr>
<tr>
<td>Spectre</td>
</tr>
<tr>
<td>Jupiter Ascending</td>
</tr>
<tr>
<td>Point Break</td>
</tr>
<tr>
<td>Avengers: Age of Ultron</td>
</tr>
<tr>
<td>Jurassic World</td>
</tr>
</tbody>
</table>
The TABLE 3 shows three hidden classes which are classified according to LFM model. From the 34 foreign films, I find that the classification sided by the producers is not so different. LFM automatic clustering algorithm classifies them into three categories. It turns out that each type is reasonable. Different types represent different groups of audience. So LFM really has the function to cluster the movies by users’ choices. And the experiments show that the accuracy of LFM is dependable compared to artificial classification.

CONCLUSION

Classification is the key in the forecast analysis. LFM automatic clustering algorithm in personalized movie recommendations from the perspective of user interest granularity diversity accordingly recommended. Although the algorithm can solve the problem of recommended in the process of particle size differences, and improve the accuracy of the recommended parameters. This article adopts manual classification results and LFM model to classify movies. Expertise is based on the subjective judgment of the LFM model based on probability for more granular clustering. When more categories appear, LFM model will show better performance. The experimental results show that the classification of LFM model has better performance than the expert classification, especially in the specified date foreign films released in China. As a future work, it is necessary to optimize the LFM model algorithm, to reduce the time complexity.

REFERENCES