RECOMMENDATION ALGORITHM BASED ON TIME CONTEXT AND TAG OPTIMIZATION

ZHANG KAI, LI FEI DA, ZHANG XU QIAN
School of Computer Science and Technology, China University of Mining and Technology, Jiangsu, Xuzhou 221116, China
E-mail: 15605205992@163.com

ABSTRACT

At present, some progress has been made in the research of context-based recommendation algorithm and label-based recommendation algorithm. However, there are some problems such as sparse scoring data for items by users and low precision of recommendation results. In response to the above problems, this paper proposes a recommendation algorithm that integrates time context and tag optimization. The recommendation algorithm is improved by integrating user behavior time interval and user attribute label information. Firstly, Long Short-Term Memory (LSTM) is introduced to study the effect of time interval on tags. Then, each output layer is combined with Latent Dirichlet Allocation (LDA) to weigh the tags with high importance. Finally, the prediction value is obtained by fusing the scoring information. Experiments show that the new algorithm has effectively alleviated the problem of sparse scoring data and improves the precision of recommendation results.

Keywords: recommender system; time context; tags; Long Short-Term Memory; topic model; Scoring Matrix

1. INTRODUCTION

With the advancement of the Internet and the development of the sharing economy, various services provided by the Internet technology have not only become more and more abundant, but their costs have also gradually decreased, bringing many convenient experiences to users' lives [1] [2]. At the same time, this makes users also troubled by how to find information that meets their needs in a mass of information, that is, the "information overload" problem. Therefore, how to alleviate this kind of "information overload" problem, assist users to find items that meet their needs, and help merchants recommend suitable items to target users, has become very important [3].

The history tag record can reflect the user's preferences, and combined with the time context information can better capture the preference rules. Through the analysis of human time behavior characteristics, it can be found that as the span of the time interval increases, the influence of its user's corresponding tag information will gradually decrease. In other words, when the time interval of the tag information is too large, it will result in low or no relevance to the upcoming event of the user. Moreover, the degree of relevance of each label to the user is different, and the importance of different labels to the user is different. Considering the impact of time interval and label importance, this paper proposes a recommendation algorithm based on time context and tag optimization to improve the accuracy of recommendation.

2. RELATED WORK

At present, some research results show that the behavior interval distribution of human behavior does not obey the Poisson distribution, but presents power-rate characteristics [4]. This means that after a relatively long period of inactivity, users will suddenly have high-frequency events in a short period of time. From this characteristic of human behavior, it can be known that certain periods of time have a great impact on users, and certain periods of time have little impact on users themselves. With the popularity of the Internet, a large amount of data including user activity time provides great convenience for better characterizing user interests and inferring user activity rules. Zhang et al. Combined the time-series correlation analysis of check-in locations and found that the time interval between similar users visiting the same place is concentrated within one hour [5]; Zhao et al. research found that the main factors affecting users to start new activities are Historical
improved. Regarding the degree of relevance of the recommendation system has also been continuously illustrated. The above analysis further explains the reason for the proposed new algorithm in this paper.

With the further development of the recommendation system, the label-based recommendation system has also been continuously improved. Regarding the degree of relevance of the tags, the term "co-occurrence" needs to be introduced here, that is, the mode of co-occurrence between tags. For example, the tags "Apple", "mobile phone", and "iPhone" often appear together, and the tags "data mining" and "machine learning" appear together. The co-occurrence of tags can reflect the association between tags. At present, the algorithm research that uses tag co-occurrence to improve the recommendation algorithm has been developed. [8] uses the item set method to calculate the correlation between tags. In [9-10], the number of co-occurrences of tags is directly used as a value for calculating the close relationship between tags. In [11], in the co-occurrence mode of tags, the tags are further subdivided into tags by using the tag probability relationship diagram. Further research on importance and correlation between tags. Among them, the degrees of the vertices in the importance correspondence graph are various. When the vertices correspond to a larger degree, it indicates that the directionality of the label is more blurred; at the same time, the correlation between the labels is also found to be asymmetric.

The user's interest preference usually does not maintain a very stable state and has frequent interactive behaviors, so the time series dynamic factor is particularly important. For dynamic interest modeling, recurrent neural networks have the advantage of better processing serialized data, and can dynamically display the temporal sequence behavior of users [12] [13]. Combining related technologies of deep learning with label recommendation can Better learn users' long-term and short-term interest preferences [14]. With the deepening of research, scholars have discovered that the gradient neural network disappears when it processes important information with long delay. In order to solve this problem, some researchers have improved the neural network through embedded methods [15] [16], derived from Long Short-Term Memory (LSTM) [17] and Gated Recurrent Unit (GRU), these methods can effectively process the user's time series data information.

Latent Dirichlet Allocation (LDA) topic regression model, the popular point interpretation is the conditional probability of words in a word list [18]. In the new algorithm proposed in this paper, all the user's existing tag information is used as the full semantic information of a topic, and LDA is used to learn the interaction between each tag and the topic, or it can be understood that under the current topic, the tag the importance of labeling and topic interaction as the attention mechanism, integrating weight values.

3. FTTO-RA ALGORITHM

This algorithm fuses the user's label information with the time interval, studies the impact of the time interval on the user's label information, and uses the LDA topic model to further improve the highly relevant label weight. According to the principle and structure of the LSTM method of the long-term and short-term memory network model, it is known that it uses the mark symbols on the conveyor belt to indicate the importance of the current information. When the information is considered important, the state is represented as 1, that is, the current information is transmitted to the next time for information fusion, otherwise throw away. Therefore, the new algorithm uses the two models of LDA and LSTM to conduct in-depth research, and proposes a Fusion Time-context and Tag Optimization Recommendation Algorithm (FTTO-RA).

3.1 Algorithm Core Structure

In order to distinguish and better integrate into the LSTM, we define a global object-tag record matrix, where matrix parameter is denoted as $U^v$, $U^v = \{(At^u_1, t^u_1), (At^u_2, t^u_2), \ldots, (At^u_m, t^u_m)\}$, $(At^u_i, t^u_i)$ represents the preference label that user $u$ has at time $t^u_i$, and $At^u_i$ represents the label. Then, the global object label record matrix is integrated into the LDA topic model to mine the specific topic probability distribution corresponding to the global object-tag. We use the project attribute matrix to represent the personalized feature label matrix information of the project itself, that is, to build the project attribute label matrix based on the project's own feature label. Corresponding to the reconstructed user tag matrix information and item tag matrix information, it is divided into two sub-
algorithms, namely user tag-based preference user set recommendation algorithm and item tag related item set recommendation algorithm. Secondly, this article introduces the LSTM model for further research, introduces the session segment into the LSTM model, and studies the impact of time interval information on user labels. (For the convenience of algorithm description, a certain time interval is referred to as a session segment in this section.). Finally, the constructed topic probability distribution is fused with the output of each session in the LSTM to further capture the more relevant labels. As shown in Figure 1.

3.2 FTTO-RA algorithm description

According to the algorithm description of the above algorithm, we integrate a certain session segment into the long-term and short-term memory network model. Here we add a specific time conversion matrix \( t_{-1} \) to store the information of session segment \( \Delta t \) and study the impact of session segment on labels. Then at time \( t \), the new candidate set is updated as:

\[
\tilde{C}_t = \tanh(W_{ch}H_{t-1} + W_u[U^u, t_{x}] + h_b)
\]

where \( W_{ch} \) and \( W_u \) are model parameters, \( U^u \) is the embedded representation of user tag information on the session segment at \( t \), \( h_{t-1} \) is the embedded representation of the output of the previous session segment, and \( h_b \) is the offset vector, and \( \Delta t \) is the session segment information at time \( t \). At the same time, in order to fully consider the impact of the session segment, we borrow time-LSTM [19] and add a time information gate \( t \) to further control the impact of the session segment on the input user tag information of the current session segment at a certain time, and finally update the status. We define the objective function as follows:

\[
C_t = f_t \cdot C_{t-1} + i_t \cdot t \cdot \tilde{C}_t
\]

\[
t_t = \sigma(W_uU^u_t + \sigma(W'_uX_t) + h_t)
\]

where \( W_u \) and \( W'_u \) are model parameters, \( U^u_t \) is the embedded representation of user tag information on the session segment at \( t \), and \( h_t \) is the offset vector, and \( \Delta t \) is the session segment information at time \( t \). In summary, we define the objective function of LSTM as follows:

\[
f_t = \sigma(W_f [h_{t-1}, U^u_t] + h_f)
\]

\[
i_t = \sigma(W_i [h_{t-1}, U^u_t] + h_i)
\]

\[
\tilde{C}_t = \tanh(W_c [h_{t-1}, U^u_t, t_{x}] + h_c)
\]

\[
C_t = f_t \cdot C_{t-1} + i_t \cdot t_t \cdot \tilde{C}_t
\]

\[
o_t = \sigma(W_o [h_{t-1}, U^u_t, t_{x}] + h_o)
\]

\[
h_t = o_t \cdot \tanh(C_t)
\]

Where \( W_f \) are model parameters, \( U^u_t \) is the embedded representation of user tag information on the session segment at \( t \), \( h_{t-1} \) is the embedded representation of the output of the previous session segment, and \( h_o \) is the offset vector.

Then we use the LDA topic model to construct the label's topic probability distribution corresponding to the label distribution. According to the prior parameters \( W \) and \( E \), LDA's joint probability distribution formula is as follows:

\[
p(\theta, \beta, z, w|\alpha, \eta) = p(\beta|\eta) \prod_{m=1}^{M} p(\omega_m|\beta_m) p(z_m|\theta) p(\theta|\alpha)
\]

We use the following formula to compute the whole label document \( \Omega \):

\[
p(\Omega|\alpha, \eta) = \prod_{i=1}^{N} \prod_{j=1}^{M_l} p(\beta_j|\eta) \sum_{m=1}^{M} \left( p(\omega_{jm}|\beta_m) p(z_m|\theta) p(\theta|\alpha) \right) d\theta
\]

Where \( N-u \) is the number of users and \( M-l \) is the number of tags per user. Then use the posterior probability to estimate \( \theta \) and \( \beta \):

\[
p(z, \theta, \beta|w, \alpha, \eta) = \frac{p(\theta, \beta, z, w|\alpha, \eta)}{p(\Omega|\alpha, \eta)}
\]
Next, we will get all \([h_1, h_2, h_3, \ldots, h_m]\) and the calculated label topic probability distribution \(\theta_u \in \mathbb{R}^{k \times 1}\) to embed each layer. The embedding method is as follows:

\[
h^u_t = a_i h_i \tag{12}
\]

Where \(a_i\) is calculated obtain the following formula:

\[
a_i = \frac{\exp(e_i)}{\sum_{m=1}^{M} \exp(e_m)} \tag{13}
\]

\[
e_i = v_u \tanh(W_c \theta_u + U_t h_t) \tag{14}
\]

Where \(v_u \in \mathbb{R}^{m 	imes m}, W_c \in \mathbb{R}^{m \times k}, U_t \in \mathbb{R}^{m \times m}\) are the model parameters, and \(e_i\) represents the current relationship score between \(\theta_u\) and \(h_i\). When the score is large, this indicates that the correlation is relatively high.

Finally, the user's label information \(U^u_h = \sum_{i=1}^{M} h^u_i\) is obtained through two embeddings of the time session segment and the topic distribution. Then it is further fully connected and integrated into collaborative filtering for recommendation. We then use \(w_{uv} = \text{sim}(U^u_h, U^v_h) + \text{sim}(R_u, R_v)\) to improve formula (15) to get the final predicted score \(r'\) for the project.

\[
r' = \bar{r}_u + \frac{\sum_{v \in S(u, K) \cap N(i)} w_{uv} (r_v - \bar{r})}{\sum_{v \in S(u, K) \cap N(i)} |w_{uv}|} \tag{15}
\]

Where \(S(u, K)\) represents the set of users who are most similar to user U's interests, \(N(i)\) is the number of users interested in item \(i\), \(r_{ui}\) and \(r_{vi}\) are the ratings of user \(u\) and \(v\) for item \(i\), \(\bar{r}_u\) and \(\bar{r}_v\) are the average scores of user \(u\) and \(v\) on all their evaluated projects.

**Table 1: Algorithm Describes the Details.**

**Algorithm:** FTTTO-RA algorithm

**Input:** User label matrix information \(U^u\) and Scoring matrix information \(R\)

**Output:** predicted ratings of item \(j\)

01: **For each** \(u \in U\) **do** // for a certain user

02: **For each** \(\Delta t_j\) **do** // time window length

03: \(t_{M_i} = \Delta t_i\) // specific time interval transition, that is, the time interval value between the user's previous and next behavior

04: **End For**

05: \(C_i = \tanh(W_c [h_{t-1}, U^u_t, t_{M_i}] + b_c)\)

// candidate state vector at time \(t\)

06: \(C_i = f_i \ast C_{i-1} + i_i \ast t_i \ast C_i\) // new state

07: \(o_i = \sigma(W_o [h_{t-1}, U^u_t, t_{M_i}] + b_o)\)

08: \(h_t = o_i \ast \tanh(C_i)\) // output at time \(t\)

09: \(h^u_t = a_i h_t\) // new output after merging topic distributions

10: \(U^u_h = \sum_{i=1}^{M} h^u_i\) // get the current user output tag information

11: \(r' \leftarrow CF(U^u_h, R)\)

12: **End For**

### 4. EXPERIMENT

#### 4.1 Dataset

In the experiment the data set used in this article is movielens-1, which obtained the updated data set in October 2016. In order to better study the effect of the fusion of tag information and time interval, we remove the users with less movie interaction events in this data, and obtain the records of interactions between 2652 users and 4,127 movies. The tag information has a corresponding time stamp. Then, data records are uniformly divided into 80% as training data and 20% as test data. In order to prove the accuracy of the experimental results, this experiment is mainly divided into two modules, which are introduced in the experimental comparison.

#### 4.2 Experimental comparison algorithm

In order to verify the effectiveness of the new recommendation algorithm proposed in this section, we have selected the following related recommendation algorithms as comparison algorithms to compare with the new algorithm proposed in this section.

User-CF: User-based collaborative filtering algorithm. The algorithm calculates user similarity based on user rating data, and then uses traditional collaborative filtering algorithms for recommendation.
LDA-userCF: A collaborative filtering recommendation algorithm for computing neighbors by combining LDA topic models. The algorithm directly converts the user-item rating information into a global object-item document, uses the LDA topic model to obtain the "hidden vector", and then uses the "hidden vector" matrix to obtain the user similarity, and then predicts the rating information.

UTLR-CF: Collaborative filtering recommendation algorithm combining label-scoring and LDA topic models. The algorithm considers three factors: label information, latent topic information hidden in the label, and scoring information, and then recommends based on the collaborative filtering algorithm. Algorithm referred to as UTLR-CF [20].

Table 2 shows the comparison of different data information used between the FTTO-RA algorithm and each algorithm.

### Table 2 Data Information Utilization

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Rating</th>
<th>Tag</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>User-CF</td>
<td>✔</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LDA-userCF</td>
<td>✔</td>
<td>✔</td>
<td></td>
</tr>
<tr>
<td>UTLR-CF</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>FTTO-RA</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
</tbody>
</table>

4.3 Metrics

The Mean Absolute Error refers to the average value of the absolute value of the deviation between the predicted score value of each individual item and the user's real score. Its main function is to reflect the accuracy of the recommendation, which is expressed as MAE. If the value of $MAE$ is smaller, the recommended accuracy is higher. The calculation formula is as follows:

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |R_{ui} - R'_{ui}|$$  \hspace{1cm} (16)

The Root Mean Square Error is the square of the deviation between the predicted score and the user's actual score, and the ratio to the number of ratings. It mainly reflects the accuracy of prediction, expressed as RMSE. The calculation formula is as follows:

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (R_{ui} - R'_{ui})^2}$$ \hspace{1cm} (17)

4.4 Experiment analysis

In this section, MAE and RMSE are used as evaluation indicators. User-CF, UTLR-CF, and LDA-userCF are compared with the FTTO-RA algorithm proposed in this paper. It is divided into two parts: one is to determine the FTTO-RA algorithm the number of user topics. The second is to compare the effectiveness of different algorithms.

1. Impact of the number of user topics in FTTO-RA

   Topic number Topic-K is mainly used to indicate the number of topics for which the global object-tag is obtained, which affects the quality of capturing user tag information and then affects the recommendation result. Therefore, the value of Topic-K affects the effectiveness of the algorithm.

   We give the average absolute error MAE and the root mean square error RMSE corresponding to the number of topics Topic-K, so as to determine the Topic-K value, so that the recommendation result of the FTTO-RA algorithm is optimal. The following experiments were performed on the above real dataset. The experimental results based on the MovieLen-1 dataset are shown in Figures 2 and Figures 3. The horizontal axis is the value of Topic-K in FTTO-RA, and the vertical axis is the average absolute error and the root mean square error, respectively. It can be seen from the trend that with the increase of the topic number of Topic-K, the value of MAE and the calculation result of RMSE both show a decline, and the trend is relatively stable in the later period. According to the evaluation index rules, it can be judged that the recommendation effect tends to increase first. After that, a stable condition appears. When the topic number of Topic-K is about 6, it is the value that M performs better. In the experimental results of R, when the topic number of Topic-K is 6, the better state appears. Therefore, in this paper, we combine the experimental results of M and R and set the topic number Topic-K to 6 to make the algorithm reach the optimal state (where User-K takes the value of 30).
(2) Comparison results

The first set of comparative experiments is as follows:

The FTTO-RA algorithm proposed in this paper is compared with User-CF and LDA-userCF, and it is verified that the improved recommendation algorithm through tagging information and labeling potential subject information can further improve the accuracy of recommendation. Considering that the recommendation result will be affected by the value of the neighbor User-K, the values of the neighbor User-K are 5, 10, 20, and 30 as the abscissas respectively, and Topic-K is set to 6. The experimental results are shown in Figures 4 and 5.

From Figures 4 and 5, we can see that both the LDA-userCF algorithm and the FTTO-RA algorithm are below the data values of the two evaluation indicators corresponding to the User-CF algorithm. Therefore, incorporating label information into the recommendation algorithm can improve the accuracy of the recommendation.

The second set of comparative experiments is as follows:

The FTTO-RA algorithm proposed in this paper fully considers the impact of time intervals on user tags. Tags with too long-time intervals will have less value. Therefore, we have performed a comparative experiment of the FTTO-RA algorithm and the UTLR-CF algorithm to prove that after merging the session segments, the recommendation results obtained by mixing multi-source information are more accurate. As shown in Figure 6 and Figure 7.
From Figures 6 and 7, we can see that, as the value of User-K in the neighborhood increases, the MAE and RMSE values of the two algorithms decrease. In the two figures, the data values of the two evaluation indexes corresponding to the FTTO-RA algorithm are below, and it is obvious that the proposed algorithm has better recommendation effect than the UTLR-CF algorithm. When the neighbor User-K is set to 5, the FTTO-RA algorithm proposed in this paper is 2.16% lower in value and 4.26% lower than the UTLR-CF algorithm. When the nearest user-K value is 30, the lowest value is obtained. The average absolute error of the FTTO-RA algorithm is 1.87% lower than the average absolute error of the UTLR-CF algorithm. At the same time, the result ratio of the root means square error the UTLR-CF algorithm is reduced by 5.78%. The second set of experiments verified the new algorithm of merging conversational segments, tag information, and scoring information is indeed conducive to improving the effectiveness of the recommendation results.

This paper comprehensively compares the experimental results. The accuracy of FTTO-RA algorithm, UTLR-CF algorithm, and LDA-userCF algorithm is better than User-CF algorithm. It can be proved that embedding tag information can improve the accuracy of recommendation. At the same time, it was found that the FTTO-RA algorithm achieved better recommendation results than the UTLR-CF algorithm, indicating that the fusion of session segment and tag information in the FTTO-RA algorithm was more helpful to improve the accuracy of the recommendation.

5. CONCLUSION

Multi-source tag information into collaborative filtering algorithm can alleviate the "sparse data" problem, and improve the recommendation accuracy. However, simply embedding it into existing recommendation algorithms cannot fully utilize the validity of the label information. Therefore, this paper uses the long-term and short-term memory network model LSTM to study the impact of time intervals on label information, combines session segments, label correlations, and scoring information, and then combines the potential topic probability distribution with the label output information of each layer. That is, the attention mechanism is introduced to capture more valuable tag information, and then it is embedded in the score information for collaborative filtering to obtain the final recommendation. Finally, we conducted experiments on MovieLens-1 data sets, and indeed proved that the new algorithm is proposed to improve the recommendation accuracy.

REFERENCES:


