A Hybrid News Recommendation Algorithm Based On K-means Clustering and Collaborative Filtering

Jing Liu, Jinbao Song*, Chen Li, Xiaoya Zhu and Ruyi Deng

State Key Laboratory of Media Convergence and Communication, Communication University of China, Beijing, China

*Corresponding author e-mail: songjinbao@cuc.edu.cn

Abstract. In the era of paper media, the information channels and the information content were integrated. With the birth of the Internet, they tended to be separated while the information channels continued to expand, which brought a massive amount of news information to process. Therefore, it’s essential for us to adopt new methods and new models to deal with all the information. This paper gives a brief overview of news recommendation technology, and proposes a hybrid news recommendation algorithm, which combines content-based recommendation algorithm and collaborative filtering, using TF-IDF method and K-means clustering technology to extract and process the features of news content, meanwhile, this paper applies SVD technology to solving the matrix sparse problem in the traditional collaborative filtering algorithm. Moreover, news popularity is taken into consideration in this paper then it combines the candidate recommendation results of each approach. At last, this algorithm achieves a better result compared to traditional recommendation algorithm’s result.

Keywords: News Recommendation, K-means Clustering, Singular Value Decomposition, Collaborative Filtering.

1 Introduction

Traditional news recommendation algorithms include content-based recommendation algorithms, collaborative filtering, and their hybrid recommendation.

Content-based news recommendation algorithm is highly interpretable since the recommendation results are completely based on user’s browsing history, but this method may also cause the homogenization problem of results [1]. The news recommendation algorithm based on collaborative filtering does not need to analyze the text content information and the hidden feature of the text [2], it’s more of a sophisticated way to filter out user’s preference with collective intelligence. As this method has diverse recommendation results, it often comes along with the data sparse problem and the cold start of new projects [3].

This paper proposes a hybrid Chinese news recommendation algorithm that combines content-based method and CF based method, then integrates multiple recommendation results and
introduces a small part of non-personalized recommendation to complete this whole hybrid recommendation algorithm.

2 A Hybrid News Recommendation Algorithm

2.1 Title Based Recommendation
Generally, news headlines often represent the high generalizations of news articles and reflect news features and user preferences, it could be used as a simple but efficient way to extract the key information of news and user.

Step 1: Segment the headlines, remove the stop words and monosyllabic words.
Step 2: Consider the remaining words as user’s keywords, count the time of each topic in the headlines that user has read as the frequency of user’s keywords.
Step 3: Return N user’s keywords which have Top-N frequency, the expression of user’s keyword is as follows:

\[ u_{\text{keywords}} = [kw_1, kw_2, ..., kw_N] \]  

\[ u(i, \text{keywords}) \] represents user i’s keywords, \( kw_j, j \in [1, N] \) represents user’s keyword j.

Indexing all the news articles on the attributes of dataset, searching all the news which headlines including user’s keywords except the ones that user has read before, then returning a certain amount of news with a specified limitation of news release time as news candidate set \( \text{Ⅰ} \).

2.2 K-Means Based Recommendation
TF-IDF is a commonly used approach to weigh the importance of each term in a document [4], the mathematical expression of TF-IDF is as follows:

\[ w(t) = tf(t, d) \times idf(t) = \frac{n_{t,d}}{\sum n_{k,d}} \times \log \left( \frac{nd}{df(t) + 1} \right) \]  

\( w(t) \) is the weight of term t in document d, \( tf(t, d) \) represents the term frequency of term t in document d, \( idf(t) \) represents the inverse document frequency, \( n_{t,d} \) is the number of occurrences of term t in document d, \( nd \) is the total number of documents, \( df(t) \) is the number of documents that include term t. This paper uses TF-IDF method to represent news with the TF-IDF value of each term in the news to extract the feature of news content as Table 1.

| News 1 | News 2 | ... | News n |
|--------|--------|-----|--------|
| term 1 | w (1,1) | w (1,2) | ... | w (1,n) |
| term 2 | w (2,1) | w (2,2) | ... | w (2,n) |
| ...    | ...    | ...  | ...    |

K-means clustering is one of the most commonly used and classic clustering methods which is also very intuitive and easy to grasp [5], the steps of this algorithm are as follows:

Criterion function:

\[ E = \sum_{i=1}^{k} \sum_{\phi \in D_i} |\phi - m_i|^2 \]  

\( m_i \) is the mean value of cluster \( D_i \).
Step 1: Randomly select K initial cluster centers.
Step 2: Allocate each sample to the nearest cluster.
Step 3: Recalculate the mean value of each cluster and update all the cluster centers.
Step 4: Repeat step 2&3 until the criterion function reaches its convergence.
Step 5: Output k text cluster centers \( C = \{y_1, y_2, ..., y_k\} \).

Through TF-IDF method and K-means method here we have content-based user’s feature as Table 2:
Table 2. User-Feature Matrix

|          | Feature_1 | Feature_2 | ... | Feature_k |
|----------|-----------|-----------|-----|-----------|
| User_1   | uf(1,1)   | uf(1,2)   | ... | uf(1,k)   |
| User_2   | uf(2,1)   | uf(2,2)   | ... | uf(2,k)   |
| ...      | ...       | ...       | ... | ...       |
| User_m   | uf(m,1)   | uf(m,2)   | ... | uf(2,k)   |

$uf(i,s)$ represents the preference degree of user $i$ for feature $s$, and Feature_1, Feature_2,..., Feature_k correspond to the column vector in C, which means the user's preference for the news in Feature_s. The calculation formula of $uf(i,s)$ is as follows:

$$uf(i,s) = \frac{\sum_{j \in \gamma_s} R(i,j)}{\sum_{j=1}^{n} R(i,j)}$$ (4)

In the formula (4), $R(i,j)$ represents whether user $i$ clicks news $j$ or not, mark $R(i,j)$ as 1 if news $j$ has read by user $i$, mark $R(i,j)$ as 0 if $j$ has not read by $i$. $news_i \in \gamma_s$ means news belongs to cluster $\gamma_s$ and thus we have the representation of user:

$$u_{(i,\text{content})} = \mathcal{C}[uf(i,1), uf(i,2), ..., uf(i,k)]^T$$ (5)

According to the cluster where each news item is located and the user's preference for the cluster s $uf(i,s)$, assign a corresponding weight value to each news, return top-N weight news with a specified limitation of news release time as news candidate set II.

2.3 Collaborative Filtering Recommendation

Singular value decomposition is a matrix decomposition technique, [6][7] which is widely used in the fields of feature extraction, data compression, etc. The input of the traditional CF based news recommendation algorithm is user-rating matrix which has very limited scalability, therefore, this paper proposes a way to replace user-rating matrix with user-item matrix as follows:

$$A_{m\times n_t} = \begin{bmatrix}
    w_{sum}(1,1) & w_{sum}(1,2) & \cdots & w_{sum}(1,n_t) \\
    w_{sum}(2,1) & w_{sum}(2,2) & \cdots & w_{sum}(2,n_t) \\
    \vdots & \vdots & \ddots & \vdots \\
    w_{sum}(m,1) & w_{sum}(m,2) & \cdots & w_{sum}(m,n_t)
\end{bmatrix}$$ (6)

In the equation (6), $m$ is the number of users, $w_{sum}(i,t)$ represents the preference degree of user $i$ for item $t$ which mentioned in section 2.2, the calculation formula of $w_{sum}(i,t)$ is as follows:

$$w_{sum}(i,t) = \sum_{j=1}^{n} R(i,j)w(t,j)$$ (7)

Normalizing matrix $A_{m\times n_t}$ by the mean value of the preference degree of user $i$ for item $t$:

$$R_{m\times n_t} = \begin{bmatrix}
    \frac{w_{sum}(1,1)}{\sum_{t=1}^{n} w_{sum}(1,t)} & \frac{w_{sum}(1,2)}{\sum_{t=1}^{n} w_{sum}(1,t)} & \cdots & \frac{w_{sum}(1,n_t)}{\sum_{t=1}^{n} w_{sum}(1,t)} \\
    \frac{w_{sum}(2,1)}{\sum_{t=1}^{n} w_{sum}(2,t)} & \frac{w_{sum}(2,2)}{\sum_{t=1}^{n} w_{sum}(2,t)} & \cdots & \frac{w_{sum}(2,n_t)}{\sum_{t=1}^{n} w_{sum}(2,t)} \\
    \vdots & \vdots & \ddots & \vdots \\
    \frac{w_{sum}(m,1)}{\sum_{t=1}^{n} w_{sum}(m,t)} & \frac{w_{sum}(m,2)}{\sum_{t=1}^{n} w_{sum}(m,t)} & \cdots & \frac{w_{sum}(m,n_t)}{\sum_{t=1}^{n} w_{sum}(m,t)}
\end{bmatrix}$$ (8)

The model description of SVD based CF news recommendation algorithm is as follows:

Step 1: Input normalized user-item matrix $R_{m\times n_t}$.
Step 2: Training SVD module
The common criteria for the assessment of SVD module are RSME and MAE [8].
Step 3: the expression of latent feature of user:

$$U_{m\times r} = [u_1, u_2, ..., u_m]^T$$ (9)
$U_{m \times r}$ is a matrix composed by the left singular vectors of $R_{m \times n_t}$ which contains the majority latent information of user after the compression of $R_{m \times n_t}$. r-dimensional vector $u_i$ represents the latent feature of user $i$.

Step 4: Calculating user similarity and predicted ratings of user for unread news with the method in [9], return top-N news with a specified limitation of news release time as candidate set III.

2.4 Hybrid Recommendation
Commonly used methods of hybrid recommendation mainly include weighting, transformation, mixing, feature combination and cascading,[10] etc. So far we have 3 candidate sets, integrating them as one and introducing a small part of non-personalized recommendation into this hybrid model to filter out the most popular news from the candidate set, which is the final recommendation result.

3 Experimental Results
The dataset used in this experiment is the news dataset published in the 2014 CFF big data competition. It includes 201M of 116,224 browsing records of 10,000 users for 6,183 news within a month. The test set is user’s last browsing record in the dataset [11]. shows the evaluation criteria used in this paper: precision, recall and F-score.

The experimental results are as Table.3 and Fig.1:

Table 3. Recommendation Results of Different Methods

| method                          | Precision | Recall | F-score |
|---------------------------------|-----------|--------|---------|
| Hybrid recommendation           | 0.8425    | 0.7671 | 0.8030  |
| Title based recommendation      | 0.4021    | 0.3520 | 0.3784  |
| Content based recommendation    | 0.7357    | 0.6791 | 0.7153  |
| Collaborative filtering         | 0.7248    | 0.6683 | 0.7032  |

Fig.1 Recommendation Results of Different Methods
Considering user_id and news_id as a set of two-dimensional data, the hybrid recommendation result can be demonstrated as **Fig.2**, which shows part of the hybrid recommendation result. The horizontal axis is user_id, the vertical axis is news_id, the red dot is the predicted result, and the blue dot is the actual result. The red dot will overlap with the blue dot and become a purple dot if the recommendation is correct.

4 Conclusion
The hybrid news recommendation algorithm proposed in this paper uses a hybrid fusion approach. The result is picked out from the final candidate set generated by a variety of recommendation methods. Candidate set I and II are both generated by content-based recommendation algorithm, using techniques such as TF-IDF and K-means clustering. Candidate set III is generated by SVD-based collaborative filtering algorithm. Then this paper integrates the multiple recommendation results and introduces a small part of non-personalized recommendation module to complete this whole hybrid recommendation algorithm and get results.

Acknowledgments
This research was supported by Social Science Research Project of the National Radio and Television Administration (No.GDT1913) and Research Project (No.HG19049).

References
[1] WANG S Q, LI X X, SUN F Z, et al. Survey of research on personalized news recommendation techniques. Journal of Frontiers of Computer Science and Technology, 2020, 14(1): 18-29.
[2] PAN Li-fang, ZHANG Da-long, LI Hui, Study on News Recommendation Arithmetic base on UserCF. Journal of Shanxi Normal University (Natural Science Edition), 2018, 32(04): 26-30.
[3] HUANG Xianying, XIONG Liyuan, LI Qindong, Personalized news recommendation technology based on improved collaborative filtering algorithm. Journal of Sichuan University, 2018, 55(01): 49-55.
[4] DAN Tangpeng, XU Tiancheng, ZHANG Shuhan. A Chinese Text Classification System Based on Improved TF-IDF Feature. Computer & Digital Engineering, 2020, 48(03): 556-560.
[5] ZHANG Guofeng, WU Guowen. Improved K-means Text Clustering Based on Kernel Function. Computer Applications and Software, 2019, 36(09): 281-284+301.

XU Zebing, WANG Zhong. Implementation of Collaborative Filtering Recommendation Algorithm Based on SVD and Hierarchical Clustering. Digital Technology & Application,
[6] LIU Qingqing, LUO Yonglong, WANG Yifei, ZHENG Xiaoyao, CHEN Wen. Hybrid Recommendation Algorithm Based On SVD Filtering. Computer Science, 2019, 46(S1): 468-472.

[7] HUANG Li, Research on Collaborative Filtering Recommendation Algorithm Based on SVD. Computer Knowledge and Technology, 2019, 15(21): 9-10.

[8] WANG Hongxia, CHEN Jian, CHENG Yanfen. Improved collaborative filtering algorithm to revise users' rating by review mining. Journal of Zhejiang University (Engineering Science), 2019, 53(03): 522-532.

[9] Zhang Zhiwei, A Survey of Personalized Recommendation Algorithms, China Computer & Communication, 2018(17): 27-29.

[10] LIU Chunxia, WU Lengmei, XIE Xiaohong. Research Review of Recommended System Assessment. Modern Computer, 2018(24): 11-15+20.