2021 International Conference on Networking, Communications and Information Technology (NetCIT) | 978-1-6654-0070-1/21/$31.00 ©2021 IEEE | DOI: 10.1109/NetCIT54147.2021.00033

A News Recommendation Algorithm Based on SVD and Improved K-means

Chang Sun
School of Computer and Information Engineering
Fuyang Normal University
Fuyang Anhui, China
sunchang2021Study@163.com

Gang Sun*
School of Computer and Information Engineering
Fuyang Normal University
Fuyang Anhui, China
ahfysungang@163.com

Zhengqi Ding
Experimental and Training Management Center
Fuyang Normal University
Fuyang Anhui, China
zhengqiding136@163.com

Qihang Liu
School of Computer and Information Engineering
Fuyang Normal University
Fuyang Anhui, China
lqh13515580125@163.com

Zhiyuan Ma
School of Computer and Information Engineering
Fuyang Normal University
Fuyang Anhui, China
MaZY15806337430@163.com

***Abstract*—With the popularity of mobile phones and other mobile devices, users can easily read news anytime and anywhere. Through the news recommendation algorithm, users can quickly find the news that matches their reading interest and improve their reading experience. How to select the type of news that users are interested in from the mass of news, this paper proposes a news recommendation algorithm based on SVD and improved K-means. This paper firstly uses TF-IDF combined with information entropy generates the news word frequency matrix, and secondly uses the result of SVD for the news-word matrix as the initial cluster center of K-means clustering, and finally uses the collaborative filtering algorithm based on users to explore the potential interests of users. Experimental results show that the algorithm has a good recommendation effect.**

Keywords—TF-ID, SVD, K-means, Collaborative Filtering, Recommendation Algorithm

I. INTRODUCTION

In this era of information explosion, users will passively obtain a large amount of redundant and useless news information. While wasting time, they may also miss news topics that they are interested in. How to quickly classify news data according to users' favorite topics. Personalized news recommendation for users is the focus of current research in the field of news recommendation [1].

Yuan Yiming et al. [2] aimed at the problem of K-means random cluster centers affecting the clustering results, using the DPMCSKM algorithm that initializes the cluster centers by density peaks. Compared with the traditional K-means method, the clustering accuracy has been improved a lot. Zhang Lin et al. [3] used the Canopy+K-means clustering algorithm for Chinese text clustering in response to the traditional K-means algorithm's inability to determine the initial cluster center and the number of clusters. The experiment proved that it is better than the traditional K-means. Li Xiaoye et al. [4] proposed an LDA topic model, combined with the K-means text clustering method, using Gibbs sampling to obtain the doc-topic matrix distribution, and then using the K-means++ method to improve the initialization of cluster centers for clustering. Xu Xintao et al. [5] first used the Doc2Vec tool to convert the news text into a vector model; then used the K-means algorithm for clustering, and finally used the improved TextRank algorithm to sort each cluster. Wen Xiaoyi et al. [6] used TF-IDF to capture keywords to generate a term weight matrix, and used SVD to decompose the term weight matrix to obtain a principal component score matrix. Zhai Donghai et al. [7] proposed the K-means text clustering algorithm to select the initial cluster center by the maximum distance method, and constructed a method to convert text similarity into text distance, and at the same time reconstructed the cluster center in the iteration. Zhao Xiaoping et al. [8] used TF-IDF to extract the feature word set of the text, and then used the Skip-gram model to get the word vector matrix through the Word2Vec, and finally used the WMD distance to calculate the similarity between the texts to achieve text clustering.

In view of the advantages and disadvantages of the above algorithms, this paper proposes a news recommendation algorithm based on SVD and improved K-means, referred to as SVD_K-means++CF. This paper firstly uses TF-IDF combined with information entropy generates the news word frequency matrix, and secondly uses the result of SVD of the news-word matrix as the initial cluster center of K-means clustering, and finally uses the collaborative filtering algorithm based on users to explore the potential interests of users. Experimental results show that the algorithm has a good recommendation effect.

II. THE FRAMEWORK OF THE SVD_K-MEANS++CF ALGORITHM

After using the word segmentation tool to remove the stop words and other tasks of the news, the weight of the words in the news is calculated through HIDF (TF-IDF combined with information entropy), the VSM word bag vector model is constructed. The news-word matrix is decomposed by SVD (Singular value decomposition), and takes the result as the initial cluster center of K-means. According to the K-means classification result, it is taken as different topics of news. The number of news users browsed under different topics is taken as users' ratings for different topics, a user-topic matrix is generated. Finally, using the user-based collaborative filtering algorithm, personalized news is recommended for users. The framework of this algorithm is shown in the Fig. 1.

978-1-6654-0070-1/21/$31.00 ©2021 IEEE
DOI 10.1109/NetCIT54147.2021.00033

A. Word processing

This algorithm uses HIDF method to select text feature words and calculate the weight of feature words.

Information entropy was proposed by Shannon, the father of information theory. The calculation formula (1) is:

$$H(X) = -\sum_{x \in X} P(x) \log p(x) \tag{1}$$

The monotonicity of information entropy can be understood as the lower the probability of a word, the greater the amount of information it carries, so the information entropy formula can be improved into the following formula (2):

$$H(t) = \frac{f_{tNew_i}}{n_{new_i}} \log\left(\frac{n_{new_i}}{f_{tNew_i}}\right) \tag{2}$$

where f_{tNew_i} represents the number of occurrences of the word t in the news New_i, n_{New_i} represents the number of words in the news New_i.

The TF-IDF formula (3) improved after combining with information entropy is:

$$HIDF_t = H(t) \times \log\left(\frac{N_D}{1 + \sum_{i=1}^{N} I(t, new_i)}\right) \tag{3}$$

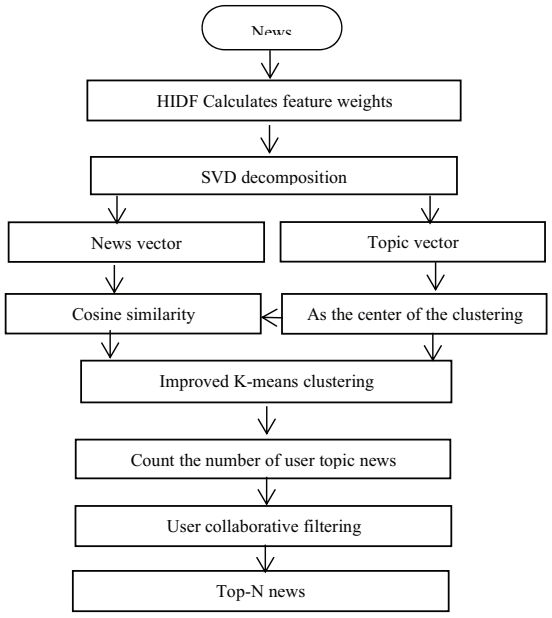

Fig.1. The framework of this algorithm

B. SVD matrix decomposition

For any m×n matrix A, then the singular value decomposition of matrix A is shown in formula (4):

$$A = U \Sigma V^T \tag{4}$$

Among them, the matrix U is an m×m left singular matrix; V is an n × n right singular matrix, U and V are real skew-symmetric matrix, satisfying $U^T U = I$, $V^T V = I$; Σ is an m × n positive semi-definite diagonal matrix, the elements $\sum_i$ on the diagonal are singular values [9] .

We can understand the matrix $\sum k$ as a certain latent semantic relationship between the news New_i and the word t_j, while U_K and V_K^T can understand a certain latent semantic relationship between the different news New_i and the difference word t_j. The relevance of the feature word t_j. The sizes of U_K and V_K^T are m×n and k×n, respectively, k is the rank of A matrix, and finally, $\sum k$ is a diagonal square matrix composed of r singular values of A (k×k) . Based on this, we can think that the singular value in $\sum k$ corresponds to the topic of the news in matrix A.

The matrix U_K can be understood as a news document-topic matrix, and each row element of the matrix U_K is the weight of a news under K different topics.

The matrix V_K^T can be understood as a topic-word matrix, and each row element of the matrix V_K^T is the corresponding word distribution and weight under a characteristic topic.

C. Similarity calculation

After the document vector is decomposed by SVD, the left matrix U_K is obtained. We take each row of topic t_j corresponding to each document d_i, which is $d_i = (t_1, t_2, ..., t_k)$. This paper uses the cosine similarity of the angle between the document vectors to calculate the document similarity. Use $\sin(d_i, d_j)$ to represent the similarity between two document vectors, which is expressed as formula (5):

$$\mathrm{sim}\left(d_i, d_j\right) = \cos\theta = \frac{d_i \cdot d_j}{|d_i| \times |d_j|} \tag{5}$$

D. Improved K-means algorithm

As a classic clustering method in machine learning and natural language processing, the K-means algorithm is characterized by its simplicity and efficiency. In addition, the clustering effect is also good in areas with larger data sets.

The basic idea of the K-means algorithm [10] is to first artificially estimate the number of clusters K. Randomly select K data from the data set as the initial cluster center of the cluster, calculate the distance from each data to the K cluster centers, assign the data to the nearest cluster according to the "proximity principle". Recalculate the cluster center until the cluster center does not change, so that the data similarity in the same cluster is large, and the data similarity between different clusters is small.

The K-means clustering method has two instabilities. One is that the system randomly initializes the clustering centers, which directly affects the final clustering effect and makes the clustering results unstable, and the other is that the choice of the number of clusters K directly affects the final effect of clustering.

The improved K-means algorithm uses SVD to matrix the news [11], and obtains the left matrix U_K, which is the incidence matrix of news-topic. Each row of U_K represents the text vector $d_i = (t_1, t_2, ..., t_k)$. Studying the topic t_j in the document d_i, you only need to set the value of other unrelated k-1 topics to 0, that is $d_{it_j} = (0,0, ..., t_j, ..., 0)$. The mean vector of these m vectors is calculated, and get $\mathrm{avg}\left(d_{t_j}\right) = (0,0, ..., \mathrm{avg}(t_j), ..., 0)$, which is the mean vector of the topic t_j. The mean vector is used as the initial cluster

center of the K-means clustering method, and K topics are decomposed by SVD as the number of K-means clusters.

The cosine similarity between document vectors is used as the K-means clustering rule for clustering, and the mean vector of all document vectors under each cluster is used as the rule for recalculating the cluster center [12], and iterating until the cluster center does not change. The iteration stops, and the K-means clustering is completed.

E. Construct rating matrix

The K-means clustering results are processed, and the number of news browsed by each user under different topics is counted as X. Perform Min-Max standardized transformation on data X, and the calculation formula is as follows (6):

$$y_i = \frac{x_i - \min_{1\le i\le n}\{x_j\}}{\max_{1\le i\le n}\{x_j\} - \min_{1\le i\le n}\{x_j\}} \tag{6}$$

For the standardized user-topic matrix, user-based collaborative filtering algorithm is used for recommendation [13]. There are many ways to calculate user similarity. Here we choose cosine similarity, as shown in the following formula (7):

$$\sin(U,V) = \frac{|N(U)\cap N(V)|}{\sqrt{|N(U)||N(V)|}} \tag{7}$$

$sim(u,v)$ is the cosine similarity of user u and user v, N(U) type user u's news topic rating set, N(V) user v's news topic rating set.

After obtaining the user similarity, calculate the user's possible score for the unrated item. According to the $sim(u,v)$ value, select the user's nearest neighbors in order, and then use the average deviation method to predict the corresponding news, as shown in the following formula (8):

$$q_{ui} = \bar{r}_u + \frac{\sum_{v\in N(u)_k} sim(u,v)(r_{vi} - \bar{r}_v)}{\sum_{v\in N(u)_k} sim(u,v)} \tag{8}$$

where q_{ui} represents the predicted rating of user u for news i, $N(u)_k$ represents the top k users with the greatest similarity to user u, and r_{vi} represents the rating of user v for news i.

F. The description of the SVD_K-means++CF algorithm

A news recommendation algorithm based on SVD and improved K-means

Input: News set S, the number of topics K, and the number of adjacent users is M.

Output: News recommendation results.

step 1: Perform news preprocessing, word segmentation, and stop word removal operations on the news set S.

step 2: The HIDF algorithm is used to calculate the weight of words in the news, and the news document-word matrix A is obtained.

step 3: Then perform SVD on the document-word matrix A, the number of singular values is k, and get U_k, $\sum k$, V_k^T.

step 4: For the document-topic matrix U_k, calculate the mean vector of the news under each topic.

step 5: Use the mean vectors of the K topics obtained in step 4 as the initial cluster centers.

step 6: Count the number of news views under K topics generated by each user after K-means clustering as X, perform Min-Max standardized transformation on data x, and map the value to [0,1] to obtain User -Topic weight matrix.

step 7: Use the cosine similarity calculation for the User-Topic weight matrix to perform user-based collaborative filtering.

step 8: Take the top M users with the greatest user similarity, perform rating prediction, and generate a Top-N news list to recommend to users.

III. Experimental results and analysis

A. Dataset

The dataset used in this experiment is the Caixin.com data set. This dataset is all news browsing records of Caixin.com users in March 2014. It contains a total of 15854 records of 3349 news items browsed by 347 users. Each record includes: user ID, news ID, reading time (time stamp), news title, news content, news category.

B. Experimental software and hardware environment

The software and hardware environment of the experiment is shown in Table I.

Table I. Software and hardware environment

Hardware environment
CPU: Intel(R) Core(TM) i7 CPU@2.90HZ
RAM：16GB
Hard disk：1T
Software environment
Operating system：Windows10 64bit
Programming language：Python3.7
Components：Numpy, jieba, Sklearn, math, matplotlib,

C. Evaluation indexes

For the news recommendation algorithm based on SVD and K-means, this paper uses three commonly used and recognized good indexes. They are Precision, Recall and F-Measure(F1).

Precision refers to the ratio of the number of hit news to the total number of recommended news, and its formula (9) is:

$$Precision = \frac{|hits_u|}{|recset_u|} \tag{9}$$

where $|hits_u|$ refers to the number of news browsed by user u in the recommended news list, and $|recset_u|$ refers to the total number of news recommended to user u.

Recall refers to the ratio of the number of hit news to the maximum number of hits, and its formula (10) is:

$$Recall = \frac{|hits_u|}{|testset_u|} \tag{10}$$

where $|testset_u|$ refers to the maximum number of relevant news that user u should recommend.

In order to effectively balance Precision and Recall, the F1 measurement standard is used, and its formula (11) is:

$$F1 = \frac{2 \times \Pr ecision \times Re call}{\Pr ecision + Re call} \tag{11}$$

The F1 value changes with the changes of Precision and Recall. When the F1 value is higher, the performance of the model is better.

D. The parameter K

Select the ideal number K to achieve the best clustering effect for the subsequent K-means algorithm. We use the Numpy component to matrix decompose the news-word matrix to obtain U, Σ, and V^T three matrices. Σ takes the middle singular value matrix to calculate the singular value percentage, and draws line graphs of singular values K and percentage, observe the change of percentage with K.

From Fig. 2, we can see that although the degree of fit is better when percentage is greater than 0.99, the K value increases and the dimension of the singular value matrix is larger. When the percentage is less than 0.99, the K value decreases and the smaller dimension is not conducive to the k-means clustering. Therefore, the percentage is selected to be equal to 0.99, and the number of K is 10.

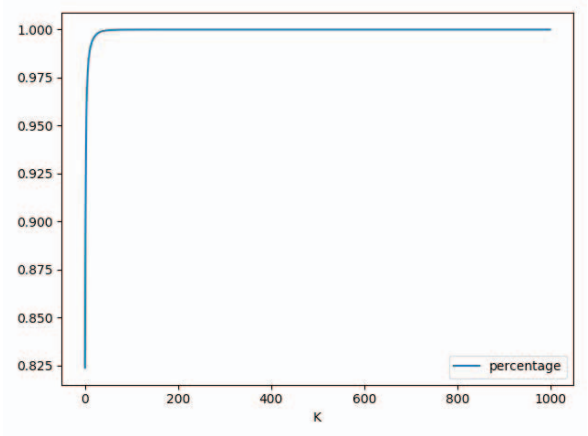

Fig. 2 The variation curve of the percentage with the K

E. Experimental analysis

In order to verify the performance of the algorithm proposed in this paper, the traditional collaborative filtering based on cosine correlation, collaborative filtering based on SVD (dimension retention 10), collaborative filtering based on K-means (number of clusters 10) and SVD_k-means. The algorithm of randomizing the initial cluster center and using Euclidean distance for clustering is compared with the algorithm proposed in this paper. On the test data set, the performance of these four methods on different numbers of neighbors is shown in the figures.

Fig. 3, Fig. 4, and Fig. 5 show that when the number of recommended news is 5, 10, 20, 25, 30, and 35 respectively, the changes in precision, recall and F1 value with the number of recommended news Top-N recommended by the SVD_K-means++CF algorithm, the traditional collaborative filtering algorithm, the collaborative filtering algorithm that only uses k-means clustering, and the SVD_K-meansCF algorithm that does not initialize cluster centers. Experimental results show that the recommendation algorithm proposed in this paper is significantly better than the traditional collaborative filtering algorithm and the collaborative filtering algorithm based on K-means, and slightly better than the SVD_K-means algorithm. It is proved that after initializing the cluster centers, the K-means algorithm clustering is more reasonable and accurate. The use of cosine similarity to calculate the news vector relationship is better than the SVD_K-means algorithm that uses Euclidean distance to grasp the internal relationship between news. Therefore, the SVD_K-means++CF algorithm proposed in this paper is better for users to make personalized news recommendations.

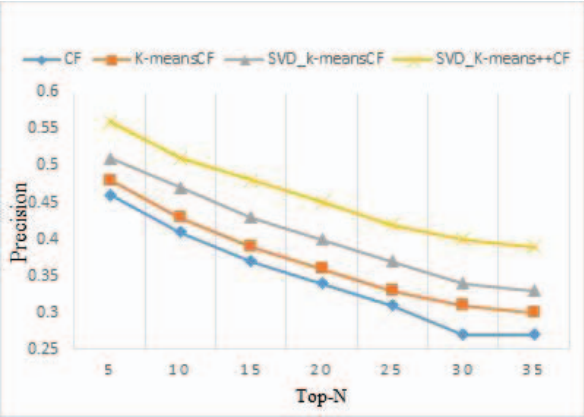

Fig. 3 The line chart of precision

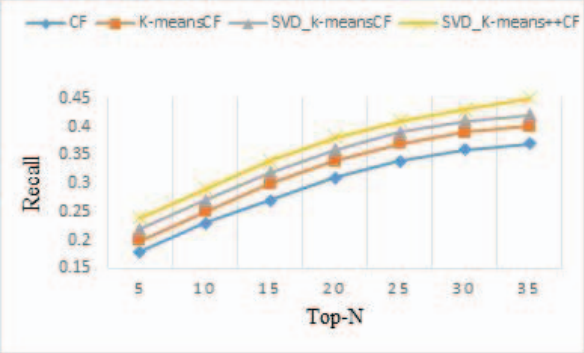

Fig. 4 The line chart of recall

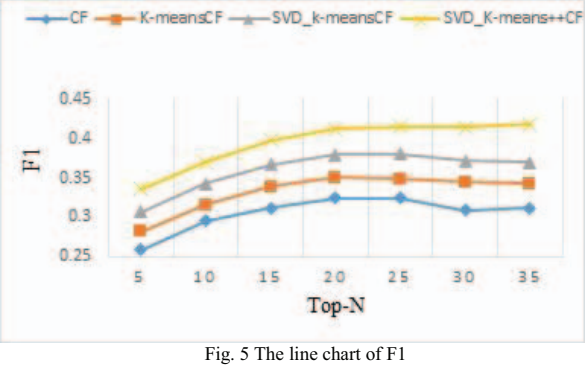

Fig. 5 The line chart of F1

IV. Conclusion

The SVD_K-means++CF algorithm proposed in this paper first performs word segmentation and stop word preprocessing on the news browsed by users, and uses TF-IDF combined with information entropy generates the news word frequency matrix. Secondly, through SVD singular value matrix decomposition, the news-topic matrix is obtained. The topic feature mean vector is used as the initial cluster center of K-means to solve the clustering instability and local optimality of the traditional K-means random cluster center. The news vector is clustered according to the cosine similarity, and then the number of users browsing news under different topics is counted to generate a user-topic matrix. Finally, it adopts user-based collaborative filtering to make personalized news recommendations for users. The experimental results verify the feasibility and effectiveness of the algorithm proposed in this paper. The next research content is to consider increasing the time factor function and the user's social relationship to make personalized news recommendations for users.

Acknowledgments

The work was supported in part by the National Natural Science Foundation of China (61906044), and in part by Natural Science Research Project of Fuyang Normal University under Grants 2020XXGN01, and supported by the Natural Science Foundation of the Anhui Higher Education Institutions of China (KJ2020ZD48, KJ2019A0532 and KJ2019A0542).

References

[1] Zhou Qingping, Tan Changgeng, Wang Hongjun. Improved KNN Text Classification Algorithm Based on clustering. Computer application research, 33 (11), pp.3374-3377,2016.

[2] Yuan Yiming, Liu Hongzhi, Li Haisheng. Improved k-means text clustering algorithm based on peak density and Its Parallelization. Journal of Wuhan University, 65 (05), pp.457-464,2019.

[3] Zhang Lin, Mou Xiangwei. Chinese text clustering algorithm based on canopy + K-means. Library forum, 38 (06), pp. 113-119,2018.

[4] Li Xiaoye, Li Chunsheng, Li long. Text clustering retrieval based on LDA model. Computer and modernization, 18(06), pp. 7-11,2018.

[5] Xu Xintao. Research on Chinese single document summary based on doc2vec and improved textrank. Electronic Science Research Institute of China Electronics Technology Corporation, 2019.

[6] Wen Xiaoyi, Hao Chengcheng. Research on news headline clustering based on singular value decomposition. Computer technology and development, 30 (02), pp. 42-46,2020.

[7] Zhai Donghai, Yu Jiang, Gao Fei. Research on K-means text clustering algorithm for selecting initial cluster center by maximum distance method. Computer application research, 31(03), pp.713-715,2014.

[8] Zhao Xiaoping, Huang Zuyuan, Huang Shifeng. A short text clustering algorithm combining TF-IDF method and word vector. Electronic design engineering, 28 (21), pp. 5-9,2020.

[9] Wang Wei, Yang Ning, Li Lihua. K-means clustering collaborative filtering algorithm based on SVD. Microcomputer information, 28 (08), pp. 139-141,2012.

[10] Guo Jinchao, Yang Jigang. Collaborative filtering algorithm based on Improved SVD algorithm and bisection K-means clustering algorithm. Journal of light industry, 35 (04), pp. 88-95,2020.

[11] Xie Z, Wang S, Chung F L. An enhanced possibilistic C-Means clustering algorithm. Soft Computing, 12(6), pp.593-611, 2008.

[12] Wang M S, Chen W C. A hybrid DWT-SVD copyright protection scheme based on k-means clustering and visual cryptography. Computer Standards & Interfaces, 31(4), pp. 757-762, 2009.

[13] Wang L, Pan C. Robust level set image segmentation via a local correntropy-based K-means clustering. Pattern Recognition, 47(5), pp. 1917–1925,2014.