Research on news recommendation algorithm based on user behavior

Li Chengcheng[a], Liu Yu[b,*], Li Zeng[c]

School of Information Science and Engineering, Guilin University of Technology, Guangxi, China

E-mail: [a]15902318506@163.com, [b]Lewis_5709@163.com, [c]597330018@qq.com

*Corresponding author

Keywords: Markoff algorithm; recommendation algorithm; accuracy; news recommendation

Abstract: In order to improve the efficiency and accuracy of the news recommendation. Through the study of user behavior and the analysis of user news browsing behavior log, a news recommendation algorithm based on Markov algorithm is adopted, with collaborative filtering algorithm and user based recommendation algorithm, and the Spark (Computing Framework) is used as the running platform, and the news recommendation algorithm based on user behavior is carried out. Research. Based on the in-depth analysis of parallel algorithm, the Mark off model is established to realize the application in intelligent news recommendation. By comparing the traditional recommendation algorithm, the test results show that the algorithm has obvious improvement in accuracy and execution efficiency, and its function is more intelligent.

1. Introduction

According to the Statistical Report on the Development of China's Internet [1], by December 2017, China's Internet users had reached 772 million and the Internet penetration rate had reached 53.2%. Among them, the proportion of mobile phone users reached 97.5%. The popularity of mobile terminals makes people more convenient and faster to get information. Intelligent and interactive network platform enables everyone to be the uploading and receiving of information. The explosive increase of information volume leads to a serious imbalance between information overload and people's demand [2]. Therefore, it is particularly important to find out the information that users want from a large amount of information and recommend it to users. The advent of recommendation system makes it easier for people to get the information they want, facilitates people at the same time, it also brings new innovation to major enterprises. Enterprises need to accurately recommend their product information to the users who need it. Users need to get the information they want. Recommendation system is a hot topic of research. Each major field and enterprise has its own recommendation system, such as food recommendation, travel recommendation, film recommendation, product recommendation, news recommendation and so on.

The emergence of intelligent news recommendation system more meet the needs of users, the system not only helps users find valuable news for users, but also can send news to the people interested in it. In personalized recommendation system, the most widely used is collaborative filtering algorithm [3]. According to the different neighborhoods, it is divided into user-based collaborative filtering algorithm and content-based collaborative filtering algorithm. Each of the two algorithms has its own problems. For user-based collaborative filtering algorithm, the algorithm itself is more likely to ignore the news itself, such as timeliness, which will lead to poor timeliness of the recommended news and reduce the actual acceptance rate of users; for content-based collaborative filtering algorithm, the user's characteristics, such as preference, are often neglected. As a result, the same kind of similar news will appear in a large number of recommendation lists, making the user news receive a fixed or smaller range.

In view of the above problems, through psychological and behavioral analysis, when people are bored with the Internet browsing information in the process, is aimless behavior. Especially news, people do not browse the news of a certain field more, but in all fields are involved. Therefore, by mining the browsing trajectory of a large number of users, it is found that whether people browse the

next news or not depends on the content quality of the current news. For example, when a particular news is read, people rarely browse the same news, because people have recently gotten the main news from the current news. On the basis of previous theoretical research[4], this is more in line with Markov decision-making, that is, the next decision only depends on the current state, and has nothing to do with the previous state. Therefore, an in-depth study of the Markov algorithm will be implemented in the news recommendation system.

2. Spark and Markoff Model

2.1 Spark computing framework

With the continuous innovation of computing framework, traditional algorithms are more or less transplanted to the new computing framework. As a memory-based computing framework, Spark is 10-100 times faster than the traditional MapReduce framework. This makes Spark a popular application.

Spark, born in UC Berkeley AMP lab, is similar to the MapReduce computing framework in Hadoop, but its performance is better than MapReduce. The computational framework is based on memory, and the intermediate result of the Spark task is[5] saved in memory compared to the result of the MapReduce calculation saved on disk. This makes it unnecessary to read and write HDFS repeatedly. Therefore, we can better use its performance to implement the iterative algorithm in big data (iterative algorithm). Its characteristics are as follows:

(1) High efficiency.The memory computing engine provides a cache mechanism to support Iterative Computing or multiple data sharing, and also reduces the IO overhead of data reading; a unique DAG (Directed Acyclic Graph) engine reduces the overhead of reading and writing from intermediate results of multiple calculations to HDFS; multilinear task pool start-up reduces shuffle Unnecessary sort operations in e process and [6-8] reduction of disk IO operations. Therefore, the computation speed is 10~100 times faster than that of MapReduce.Wherever Times is specified, Times Roman or Times New Roman may be used. If neither is available on your word processor, please use the font closest in appearance to Times. Avoid using bit-mapped fonts if possible. True-Type 1 or Open Type fonts are preferred. Please embed symbol fonts, as well, for math, etc.

(2) Easy to use.Spark provides a rich API that supports four languages, Java, Scala, Python, and R. In addition, operating on Spark requires two to five times less code than MapReduce.

(3) Currency.Because spark is integrated with Hadoop, it can be used seamlessly with many tools on Hadoop, making it convenient for developers.

The running process of Spark is shown in Figure 1. The whole process consists of Spark Context, Cluster Manager, Work Node, and Executor. Data processing logic written by Driver users; Cluster Manger is responsible for central deployment and resource management; Work Node is responsible for handling Driver commands and reporting status to Driver; Task is a computing unit; Executor is an executor. Where the Spark program runs, it creates an interactive interface for the Spark Context, schedules resources through the Cluster Manager, enables computing nodes to obtain resources, and eventually executes tasks in the Executor.

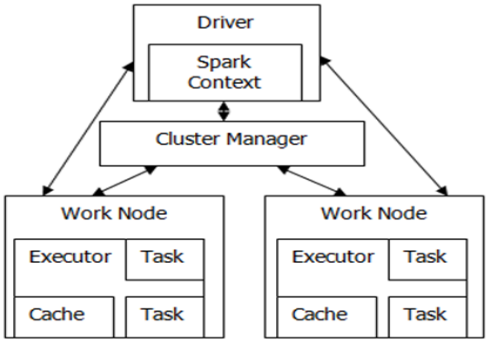

Fig. 1 The Spark program runs the flowchart

2.2 Markoff model

Markov chain,Markov chain, also known as discrete-time Markov chain (DTMC), means the process of state space from one state to another, which is "memoryless", that is, the probability distribution of the next state can only be determined by the current state state [9-10].

Described in formal languages as follows:

When the conditional probabilities on both sides of the equation are meaningful, formula 1

$$\begin{aligned} &P(X_{n+m}=j|X_n=i_n,\ X_{n-1}=i_{n-1},\cdots X_1=i_1) \\ &=P(X_{n+m}=j|X_n=i_1) \end{aligned} \tag{1}$$

When m = 1, the equation was established, and the sequence of random variables was a Markov chain. The countable set becomes the state space of Markov chain. The greatest characteristic of the chain is no aftereffect, that is, the future state of things, only related to the status of the status quo, with no previous state. This is more consistent with people's usual habit of reading news in the prediction of news recommendation.

From the user's browsing log analysis, you can visually see the news content that the user is browsing, and model its behavior and content here, as shown in Fig.2:

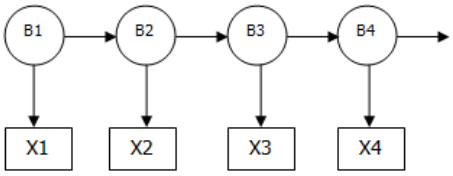

Fig. 2 Markov chain model

Among them B1, B2, B3, B4 is the user's behavior operation, that is, to click on a certain type of news page, the behavior is invisible, so it is called hidden Markov chain. X1, X2, X2, X3, X4 are the news content that users browse under this category. It is easy to find X1, X2, X3, X4 through log. Such sequences, and how to solve the probability of the next behavior, is the core of the recommendation algorithm.

The algorithm model is shown in formula 2, $P(B_{n-1}\rightarrow B_n)$ It is the probability of transformation of behavior. $P(X_n)$ becomes the probability of users browsing the next news.In order to facilitate calculation to obtain the maximum probability, this paper uses the parameter n=3 as the calculation.

$$\begin{aligned} &P(Bn)= \\ &P(B_1\rightarrow B_2)*P(X_1)* \\ &P(B_1\rightarrow B_2)*P(X_2)* \\ &\cdots * \\ &P(B_{n-1}\rightarrow B_n)*P(X_n) \end{aligned} \tag{2}$$

2.3 Related auxiliary recommendation algorithm

Collaborative filtering algorithm scores according to similar users' browsing news and forms a scoring matrix. Here we use the modified cosine formula to calculate the similarity of users. Formula 3 is as follows:

$$\mathrm{Sim}(i,j)=\frac{\sum_{u\in U}(R_{u,i}-R_u)(R_{u,j}-R_u)}{\sqrt{\sum_{u\in U}(R_{u,i}-R_u)^2}\sqrt{\sum_{u\in U}(R_{u,j}-R_u)^2}} \tag{3}$$

Among them, R_u shows the average score of user U to browse news. According to the score matrix, the top 50 news items with the highest score can be calculated and put into the recommendation list. This algorithm is used to calculate the proportion of news items with the same news.

Top N algorithm can sort the news according to the key words, at the same time, it can filter out

the news with strong relevance to the keywords, and it can also adjust the keywords with high matching. Through the top N algorithm, the recommendation of hot news can effectively solve the user's cold start problem, thus improving the accuracy of the algorithm.

If a Markov chain is set up for each user, the amount of computation is much more than the computing power. In order to reduce the computation of Markov chain and make it better used in news recommendation algorithm, we need to calculate the similarity of users.Similar users'calculation should be based on the similarity of browsing news. In this paper, we calculate similar users by setting the similarity threshold of news.

Jaccard similarity coefficient is used to calculate the similarity of users. Formula 4 shows:

$$User\_similar = \frac{UserA_{news} \cap UserB_{news}}{UserA_{news} \cup UserB_{news}} \tag{4}$$

Since the algorithm can only get the same, it is more in line with the requirements of this article. The number of news browsed by user A and user B is intersected to produce a common news set. By combining the number of user A and user B, the total number of news browsed by two users is obtained. It is the similarity between the two sets of news sets divided by the total number of news reports. Here we set a range of values, that is, $User\_similar$ is greater than 50, you can say that user A and user B are similar users.

3. Design and implementation of Markoff algorithm

3.1 Markoff algorithm state transition

Through the collaborative filtering algorithm, we can easily calculate the percentage of users browsing each type of news, that is, probability P(B). In each category of news, the number of news voted out can be used as an optional number of such news, if the first five of each category of news as a candidate, then the probability of these five news is 1/5.

In practice, the number of news readers per user is limited, which leads to the different length of Markov chain. According to the above models, we use the latest three user browsing records as the parameter n of the model. Calculate the maximum probability of the next action through three recent records.

This paper chooses the last three operations as B1, B2, B3, and solves B4 as the prediction user's click action. The state transition process is as follows:

$$P(B1) = \begin{bmatrix} P(X1) \\ P(X2) \\ P(X3) \end{bmatrix} \tag{5}$$

Under the action of B1, the probability of user clicking can be P (Xi), i=1, 2, 3.

$$\begin{aligned} P(B_2) &= P(B_1 \to B_2) \\ &= \begin{bmatrix} P(X_1) \\ P(X_2) \\ P(X_3) \end{bmatrix} * \left[P(Y_1)P(Y_2)P(Y_3)\right] * P(B_2) \\ &\begin{bmatrix} P(X_1Y_1)P(X_1Y_2)P(X_1Y_3) \\ P(X_2Y_1)P(X_2Y_2)P(X_2Y_3) \\ P(X_3Y_1)P(X_3Y_2)P(X_3Y_3) \end{bmatrix} * P(B_2) \end{aligned} \tag{6}$$

Formula 6 is the transition process of $B_1 \to B_2$ state, where Y_i represents the content of the second kind of news, and $P(Y_i)$ is the probability of that kind of news. Therefore, the probability of B2 may be derived. Similarly, the state of B is similar, so the probability of B4 existence can be obtained.

$$P(B_3 \to B_4) = P(B_3) * [P(C_1)P(C_1)P(C_1)] * P(B_4) \quad (7)$$

Select the top news of maximum probability as recommended list. This can satisfy the recommendation of Markov algorithm.

3.2 Implementation of Markoff algorithm

In this paper, Markoff algorithm is used as the main algorithm, combined with collaborative filtering algorithm and Top N algorithm. Collaborative filtering algorithm can quickly score news, and Top N algorithm can quickly select the best news into the candidate list of Markov algorithm. Therefore, these two algorithms are more likely to get candidate news sets than traditional sorting lookup. Wayne diagram is shown in Fig. 3.

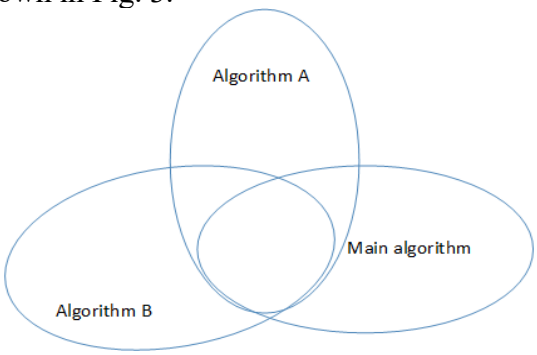

Fig. 3. Result set Wayne diagram

Algorithm A is a content-based collaborative filtering algorithm, and algorithm B is a Top N algorithm. The overlapping parts of the two algorithms and the set from the main algorithm are used as the recommendation list. This eliminates the excessive computation of the main algorithm and increases the accuracy of its recommendation.

Through the above combination algorithm, the flow chart of the experiment in the Spark framework is as follows:

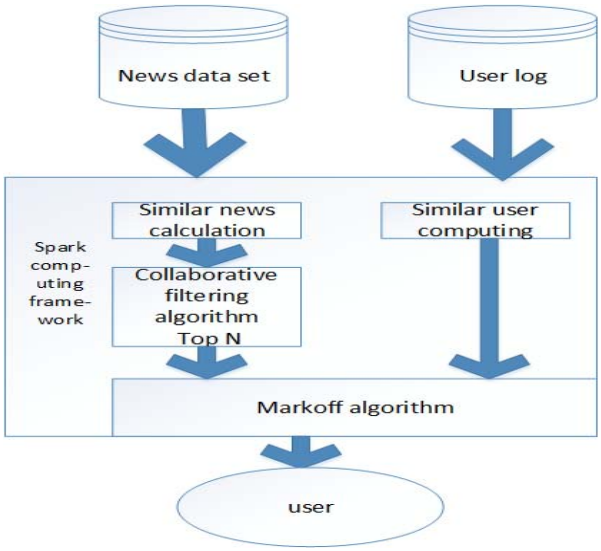

Fig. 4. Algorithm implementation flow chart

The experimental procedures are as follows:

Step 1: According to the news data set and user behavior log, similar news and similar users are calculated respectively, and similar news and similar users are grouped into the same set.

Step 2: The test set of similar news is used to collaborative filtering algorithm, and the recommendation list is obtained by collaborative filtering algorithm. At the same time, the top three news with the highest clicks in each similar news set are calculated by Top N algorithm in parallel mode, which is used as the Markov calculation list.

Step 3: two sets of recommended lists are collected and intersected.

Step 4: According to the Markov algorithm, a user is randomly selected from a set of similar users as the basis of calculation, and the news ID of the last three browses of the user is calculated. The next news of all the browsing users of the three news is counted and the Markov chain is generated.

Step 5: if the news exists on the union, enter the new recommendation list, if not, then abandon it. Combine the newly generated recommendation list with the news on the intersection to get the final recommendation list. According to the Markov chain, the top 30 news with the highest number of votes are selected and pushed to the user.

The pseudo code of its Markov algorithm is as follows:

```
begin
initialize β,(T), t ← T, a_{i,j}, b_{j,k}, V^T
for t ← t − 1;
β_i(t) ← Σ_{j-1}^{i} β_j(t + 1)a_{i,j}b_{j,k}v(t + 1)
until t = 1;
return P(V^T) ← β_i(0),
end
```

Among them, T is a collection of recommended lists overlapped by collaborative filtering and Top N algorithm. The visible sequence is the user behavior sequence, that is, users browse the news sequence. Set for similar users.

4. Experiment and result analysis

The experimental data set in this paper is the behavior log of 50,000 users crawled from Sina News website and the news set they browsed. The hardware experimental platform selected in this paper is a number of the same models of computers, the operating system is CentOS 6.5, Spark version is 2.2.0, the processor is Intel I7 Quad core processor, 8G running memory, 1T hard disk.

This paper mainly realizes the application of Markov algorithm based on Spark in recommendation algorithm. By supplementing the content-based collaborative filtering algorithm and Top N algorithm, it has better accuracy. In this paper, two experiments are conducted to verify the superiority of the algorithm.

The recall and accuracy are used to evaluate the results of the experiment. The formula is as follows:

$$precision = \frac{\sum_{u_i \in U} hit\ (U_i)}{\sum_{u_i \in U} R\,(U_i)} \tag{8}$$

$$\text{recall} = \frac{\sum_{u_i \in U} hit\ (U_i)}{\sum_{u_i \in U} R\,(U_i)} \tag{9}$$

Here we use the user set as a recommendation to indicate the number of users browsing news in the test set. In this experiment, each user has a record, so its value is 0 or 1. Set represents the total number of recommended news.

The accuracy and recall of Markoff algorithm in the recommendation process are calculated. By dividing the data into five different data quantities and calculating the accuracy and greeting rate of different data, it is shown in Fig. 5:

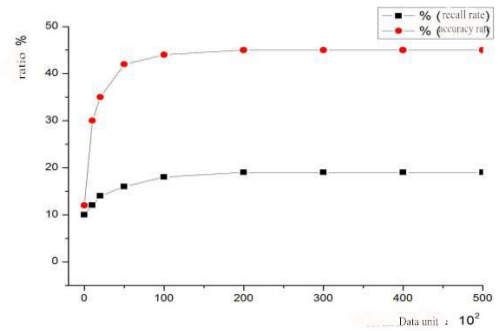

Fig. 5. Accuracy rate and recall rate calculation

As can be seen from Fig. 5, the performance of the algorithm is related to the complexity of the data. As the amount of data increases, more and more "outdated" news will appear in the news set. For Markov algorithm, the performance of the algorithm increases with the number of recommendations. When the amount of data is above 10000, its accuracy is stable at 45%, while the greeting rate is stable at 19%. That is, when the threshold exceeds a certain threshold, the algorithm will neither improve nor reduce performance.

Through experiments on different algorithms, the accuracy and recall rates of different algorithms are obtained under the same data. By comparison, the results are shown in Fig.6:

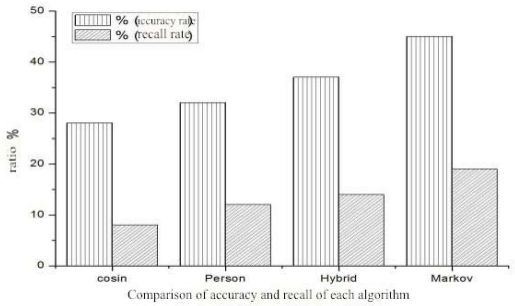

Fig. 6. Comparison of accuracy and recall of each algorithm

Fig.6 shows that under the same data, the accuracy of collaborative filtering algorithm based on cosine similarity is 28% that of Person-based collaborative filtering algorithm is about 33%, that of content-based recommendation algorithm is 38%, and that of Markov-based algorithm is 45%, which shows that the accuracy of this paper is based on Markov. The application of Koff algorithm in news recommendation system is obviously superior to other recommendation algorithms, and the recall rate and accuracy rate are obviously improved.

Experiments show that Markov-based algorithm combined with its content-based collaborative filtering algorithm and Top N algorithm has a higher accuracy in news recommendation than the traditional algorithm.

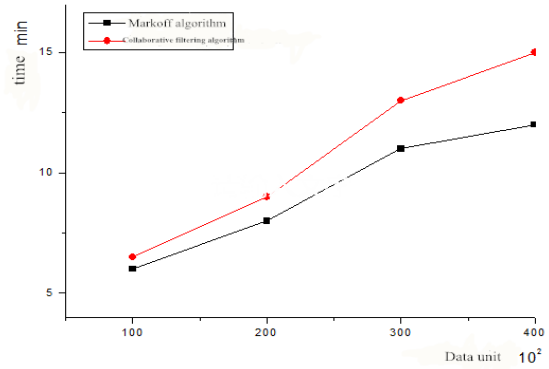

Fig. 7. Algorithm running time comparison

By comparing the running time of the content-based collaborative filtering algorithm in the same data, its running efficiency is verified. The experimental results are shown in Fig.7:

As can be seen from Fig. 7, the running time of the Markov algorithm is relatively less than that of the content-based collaborative filtering algorithm under the same operating environment and the same operation data, that is, the running efficiency of the Markov algorithm is better than that of the collaborative filtering algorithm. However, due to the input-output delay of the hybrid algorithm, the effect is not too obvious.

5. Conlusion

Starting from people's behavior of browsing news, this paper analyzes the purposelessness of browsing news, which is closer to Markov chain. Therefore, the Markov algorithm is applied to the news recommendation system. A large number of experimental data show that the proposed algorithm is more effective than the traditional recommendation algorithm in the recommendation process. Its accuracy and operation efficiency are better. However, the proposed algorithm has some limitations, that is, the computational complexity is determined by the matrix formed by the selected behavior parameters. If the behavior parameters are too large, the amount of data will increase, and the computational complexity will increase accordingly, thus affecting the recommendation efficiency. Therefore, in the follow-up study, we will further optimize the efficiency of the algorithm and the choice of behavior parameters to improve the accuracy of its recommendation.

References

[1] Statistical report on China's Internet development [R].China Internet Information Center.2018.1.18.http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/hlwtjbh/201803/P020180305409870339136.pdf.

[2] Liu Can, Ren Jianyu, Li Wei. Personalized Recommendation Oriented Educational News Crawling and Display System [J]. Software Engineering, 2018 (02): 38-40+34.

[3] Li Letian, Wu Lin. research on news recommendation algorithm [J]. Journal of Communication University of China.2016.2.

[4] ONNALAGEDDA N, GAUCH S. Personalized news recommendation using Twitter [C] / / WI-IAT 2013: Proceedings of the 2013 IEEE /WIC /ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies. Washington, DC:IEEE Computer Society, 2013, 3: 21-25

[5] Qie Zhi-hao. Research and implementation of news recommendation system based on Hadoop [D]. South China University of Technology, 2016.

[6] J. Dai, “Experience and lessons learned for large-scale graph analysis using GraphX,” 2015, Spark Summit East.[Online]. Available: https://spark-summit.org/east-2015/talk/experienceand-lessons-learned-for-large-scale-graph-analysis-using-graphx.

[7] J. Jiang, J. Lu, G. Zhang, and G. Long, “Scaling-up item-based collaborative filtering recommendation algorithm based on hadoop,” in Services (SERVICES), 2011 IEEE World Congress on. IEEE, 2011,pp. 490–497.

[8] Wang Zikun's translation. Markov process theory [M]. Harbin Institute of Technology press. 2015.

[9] Sun Xiaohui. Personalized news recommendation system based on user behavior [D]. University of Electronic Science and technology, 2015.

[10] ZUO Y C, YOU F, WANG J M, et al. User modeling driven news filtering algorithm for microblog service in China [C]// ICIS '12:Proceedings of the 2012 IEEE /ACIS 11th International Conference on Computer and Information Science. Washington, DC: IEEE Computer Society, 2012: 393 -399