Available online at www.ijpe-online.com

International Journal of Performability Engineering

vol. 14, no. 5, May 2018, pp. 1078-1087
DOI: 10.23940/ijpe.18.05.p27.10781087

A Mining Model of Network Log Data based on Hadoop

Yun Wu[a], Xin Ma[a], Guangqian Kong[a,*], Bin Wang[b,c], and Xinwei Niu[c]

[a]*College of Computer Science and Technology, Guizhou University, Guiyang,* 550025, *China*
[b]*School of Mechanical Engineering, Yancheng Institute of Technology, Yancheng, 224000, China*
[c]*School of Engineering, Penn State Behrend, Erie, Pennsylvania, 19019, United States*

Abstract

With the increasing amount of data in the information age, traditional Web log data mining method has been unable to deal with large-scale text data. Aiming at these problems, we design a high reliability Web log data mining scheme and put forward a kind of text similarity simulation detection model based on Hadoop. Firstly, we design a data mining scheme for user behavior log, which considering the heterogeneity, diversity and complexity of network log data. The design of the platform is divided into three layers: Data storage layer, Business logic layer, and Application layer. In this part, we design the data cleaning algorithm and KPI, and then use Hive to complete mining. Secondly, a text log data similarity mining model based on Hadoop is proposed, and the algorithm of text similarity mining model is designed. This mining model including the Shingling algorithm and NewMinhash algorithm for the design of MapReduce. Using the improved Shingling algorithm based on the MapReduce programming model, the document is converted to a collection. The distributed New Minhash algorithm is used to solve the signature matrix, and the Jaccard coefficients are used to calculate the similarity. We conduct experimental analysis based on data set SogouCS. The experimental results show the effectiveness of the NewMinhash algorithm, and prove that the model can not only find the similarity of text accurately, but also can better adapt to the distributed platform, and have good expansibility.

Keywords: network log; data mining; Shingling algorithm; NewMinhash algorithm; Hadoop platform

(Submitted on January 25, 2018; Revised on March 13, 2018; Accepted on April 17, 2018)

1. Introduction

With the rapid development of the Internet, the number of log data has increased exponentially [7]. The huge amount of data and the limitation of technology and tools is not easy to analyze [3]. We mainly research on user behavior log data and text log data. Web log mining technology has been widely used in the analysis of web log data. The traditional data mining technology is inefficient in data analysis of massive network log data, which requires the application of distributed technology to deal with large scale computing and distributed storage by using the MapReduce programming model in Hadoop. Moreover, the huge network log data is often very complex, and there are errors and inconsistency in the data, which need to be eliminated or corrected [6,8]. Only washed data can guarantee the high quality of data. Because of the large amount of data and the inefficient and time-consuming of the traditional data cleaning method, it is no longer used in the processing of massive text [2,11]. Therefore, this paper mainly studies the use of Hadoop platform for data cleaning.

Based on the large number of log texts and complex analysis, this paper proposes a scheme of web log data mining based on Hadoop. We research include two main part. Through the study of Hadoop ecosystem and data mining technology, this paper firstly analyzes the data mining of user behavior log data, and puts forward a solution to mining web log data. We design a high reliable network log data mining frame, and we realize the distributed cleaning filtering algorithm based on MapReduce. We use this algorithm to filter out the high-quality data set, and provide to the data mining platform. We use this frame to dig out valuable information to support website decisions. Then, a large-scale text similarity solution based on Hadoop ecosystem is proposed for text log data. The traditional text similarity algorithm is improved to adapt to the distributed platform. We mine similar documents in the network log data, which provides a reference for solving the

* Corresponding author.
E-mail address: gq_kong@163.com.

problem of homologous press releases and plagiarism pages in the distributed environment. The second part deals with text log data; we present a solution of large - scale text similarity based on Hadoop ecosystem, which improves the traditional text similarity algorithm to adapt to the distributed platform.

This paper consists of the following sections. The first part is the introduction. The second part is the introduction of related technology and theory; in the third part, the design framework of user behavior log data mining scheme is proposed. The fourth part includes the proposed similarity mining model of text log data and the experimental results; the fifth part is the summary of this paper and the application prospect of this scheme in the future.

2. Related Work

User behavior log data is one of the network log data, which contains a lot of valuable information. By analyzing user behavior log data, it is helpful for enterprises to master user behavior patterns, understand users' interests, requirements, capabilities and other information, and provide decision support for enterprises, providing users with better services [13]. The text log data in the web log data is the log data used to record the content of the web page. This type of log can be used to analyze the homologous press release of the website, copy the web page, or copy the text nearly repeatedly.

The traditional user behavior log data mining is mostly based on relational database, so it depends on the high-performance machine to a great extent; the mining time is long and the efficiency is low. These problems will seriously affect the business decisions of enterprises. In recent decades, a large number of domestic scholars have carried out research on the mining of web log data based on Hadoop. In 2010, Hu et al. [4] made a deep research on the network log analysis system based on Hadoop, and realized the simple analysis of the network log under the Hadoop platform. An analysis system is proposed, but the analysis function is too simple and the mining depth is shallow. Yang et al. [12] made a deep research on the online network log analysis system based on Hadoop, and realized the function of mass data hierarchical optimization storage. However, the system framework is complex, and the parallel processing of data is not suitable for the system. Li et al. [5] made a deep discussion on the network log analysis based on data mining, and through the comparison of patterns and the application of data mining technology, realized the timely discovery of illegal behavior or abnormal operation of internal users. Furthermore, the security policy can be adjusted in time to ensure the security of the system. But the drawback is that we can't use efficient data mining method to reflect the behavior of users' special attributes according to the actual requirements. This paper proposes a Hadoop based network log data mining scheme using cheap computer clusters to design a user behavior log data mining scheme to solve these problems, and designs and builds a set of stable clusters as an experimental environment.

For text log data similarity mining, the commonly processed methods are divided into two kinds. The first method is based on the extension method, which does not need to care about the semantics of the text, only according to the external characteristics of the text to calculate the similarity. The big problem with this approach is that it can't be extended to large-scale text processing [10,14]. The second method is based on semantics. It is necessary to understand the semantics of text before it can be processed. Due to the restriction of natural language processing technology and the low degree of parallelization, it is not suitable for distributed environment processing [1,9]. We improve the algorithm of the traditional text similarity model to adapt to the Hadoop platform, and propose a text log data mining model based on Hadoop cluster technology, mining the similarity between text Log data. A new distributed NewMinhash algorithm is proposed, which can adapt to the distributed platform. Experiments are used to prove the correctness of the algorithm, the adaptability of the distributed platform and its expansibility.

3. Design of user behavior log data mining scheme

At present, many websites produce index log data every day. Because these data are a kind of "dark data", many enterprises do not pay attention to it, which has caused many valuable information to be buried. Moreover, many of the current data mining methods are based on traditional relational databases, and mining time is very long and efficiency is under the influence of business decisions. In this paper, we design a mining scheme based on web log data to solve these problems by using cheap computer cluster.

3.1. Design of framework

The design of the platform is divided into three layers, as shown in Figure 1, by combining these characteristics and programming ideas of the network log data with the characteristics of heterogeneity, diversity, complexity and so on.

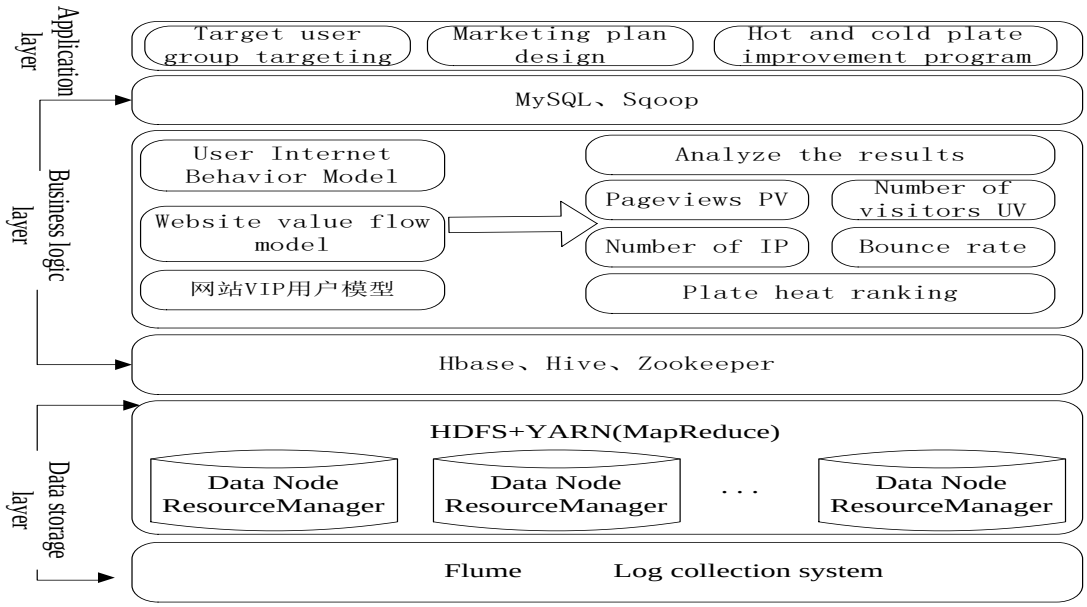

Figure 1. Design framework

Data storage layer: includes log collection system and storage calculation of underlying data. The log collection system is mainly responsible for the collection of web logs, and then calls the filtering algorithm to filter and file it to the Data Node bottom data node for storage. All the results are also stored in the data storage layer. It includes cleaned intermediate data, data mining results, log metadata, etc.

Business logic layer: core of the platform. The tools, HBase, Hive, Zookeeper, not only implement the encapsulation of business logic, but also simplify the use of users, so these tools are used for data processing and data mining. Then, according to user behavior model, website traffic model and VIP user model, valuable information such as pvp IP number can be mined. The analysis results are stored in the form of files on HDFS. The metadata of Hive uses MySQL as the storage engine through the user interface CLI (command line interface). Finally, we use the tool, Sqoop, to export the analysis results to the relational database MySQL outside the cluster.

Application layer: uses the analysis results in the business logic layer such as VIP users, block heat, page views and other user characteristics and traffic statistics results. We can locate the target user group and design the accurate marketing scheme. We can also improve the web pages, improve the user experience of the site and other decision-making services

3.2. Analysis and Mining of Network Log Data

Firstly, flume is used to simulate the generation and capture of logs, which, like the real production environment, ensures the authenticity of data collection. Then, the designed cleaning algorithm based on MapReduce is used to clean and filter the data. Designing KPI (Key Performance Indicator), use Hive for multidimensional analysis. Finally, the Sqoop tools were used to import the results into MySQL for visual display. The system administrator manages and monitors the cluster, and users mine the information by the Web UI.

3.2.1. Designing data cleaning algorithm and KPI

(1) Design of cleaning algorithm
The log data used in the experiment is a set of open data for a website. The format of the original data is shown in Table 1.

Table 1. Raw data format

Log data format	meaning
110.52.250.143	User IP
[30/May/2014:17:38:20...	Request date
GET、POST...	Request mode
/static/image...	Access resources
100-599	Access status
1127	Access traffic

A total of 540,000 rows of data, log data each row is composed of six parts: user IPs, request date, request mode, access resources, access status, and access traffic. Firstly, the data is uploaded to HDFS. Then, the filtering algorithm is designed according to the data format. From Table 1, we can see that the web site's log data format is a bit messy and needs to be cleaned and filtered for data. There are many ways to request HTTP pages, and only the method of GET can represent the user's behavior intention, so it is necessary to eliminate other ways. Then, only the beginning of 2 indicates that the user access is successful, and other noise status codes need to be removed. The algorithm is implemented by Java, and the algorithm flow is shown in Figure 2.

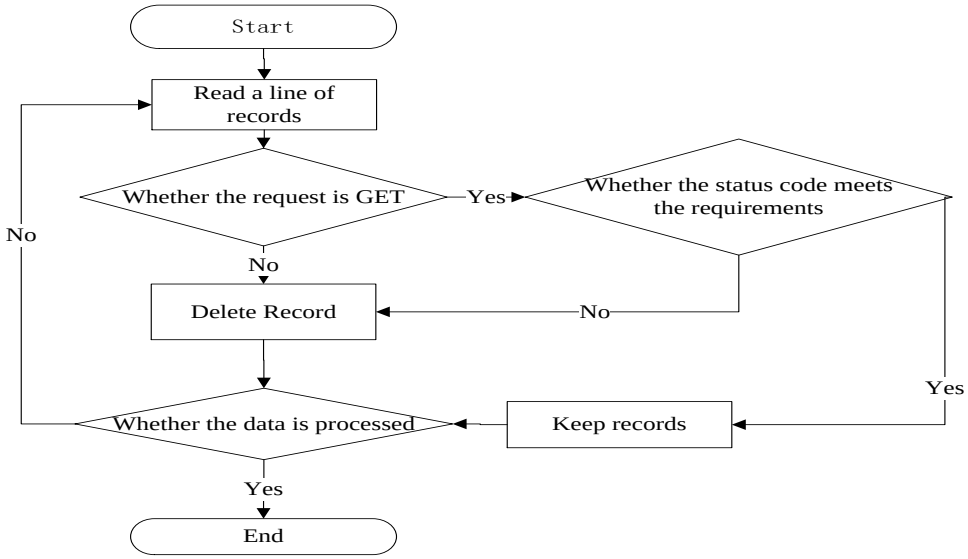

Figure 2. Filter algorithm flow

(2) Design of KPI

One of the biggest differences between data mining and data analysis is that the target information of data mining is not clear, and the techniques and algorithms used in data mining are also uncertain. Therefore, this paper specifies the design of KPI according to these key indicators.

①page views PV (Page View)

Definition: The number of times a page is browsed by a visitor, and each individual user browses a page that will add one to the items [15].

Analysis: A web site is made up of a page. The sum of all page browsing is the amount of web browsing, reflecting the customer's interest in the site [15].

Formula: Number of records.

②P number

Definition: The sun of the number of different IP sites visited over a period of time.

Analysis: In general, the number of IP is directly proportional to the access of the web site.

Formula: Count different IP.

③ounce rate

Definition: The percentage of visitors who visit a page of the site and then leave as a percentage of the total number of visitors [14].

Analysis: The bounce rate can measure the attractiveness of the website to the user and is inversely proportional.

Formula: Count IPs with only one record in a day, known as the jump out number T

Bounce rate = T / PV * 100%

④section heat ranking

Definition: Page access ranking.

Analysis: In any case, whether users like a section of the web page, it can help enterprises consolidate hot plate, strengthen the construction of cold plate.

Formula: Sort by number of visits and residence time.

3.2.2. Use Hive for data mining

(1) Initialization

Create an external table under hive to connect the data on the HDFS and store the external source data in the MySQL.

(2) Write shell scripts

HDFS log data is processed once a day, filtered with MapReduce and written to a shell daily.sh.

(3) Analyze the top three users of traffic

Enter the query command in Hive and import a table named PV, as shown in Figure 3.

```
hive> select * from hmbs limit 3;
OK
110.52.250.126  20130530173820  data/cache/style_1_widthauto.css?y7a    20150825
110.52.250.126  20130530173820  source/plugin/wsh_wx/img/wsh_zk.css     20150825
110.52.250.126  20130530173820  data/cache/style_1_forum_index.css?y7a  20150825
Time taken: 4.966 seconds, Fetched: 3 row(s)
hive> select * from hmbbs limit 3;
```

Figure 3. Top three users

Figure 3 shows the top three users of traffic on a certain day, manifesting that these users are not only the old customers who often visit the website, but also the customers who contribute a lot of traffic. They can make personalized marketing recommendations for these old customers.

(4) Query the top 10 users

Enter the query click command in Hive, as shown in Table 2.

Table 2. Traffic ranking top ten users

User IP	click rate
61.50.141.7	4855
222.133.189.179	3942
60.10.5.65	1889
220.181.89.156	1877
123.147.245.79	1571
61.135.249.210	1378
49.72.74.77	1160
180.173.113.181	969
122.70.237.247	805
125.45.155.27	735

These users in Table 2 can be considered as VIP user groups and focused on these user groups.

(1) Query page views and import them into mysql using Sqoop, as shown in Figure 4.

ip	acc_time
49.72.74.77	1160
180.173.113.181	969
61.50.141.7	4855
222.133.189.179	3942
60.10.5.65	1889
122.70.237.247	805
125.45.155.27	735
220.181.89.156	1877
123.147.245.79	1571
61.135.249.210	1378

Figure 4. Page views rank

From the ranking, we can see which pages are higher popular or lower popular with the user, provided to the enterprise for reference.

4. The Mining Model of Data Log Similarity

4.1. The method of text similarity mining model

With the amount of data increasing, the traditional mining model cannot detect the similar text quickly and accurately. Therefore, a text similarity mining model based on Hadoop distributed cluster technology is proposed in this paper. The detection process for the model is shown in Figure 5. The massive documents must be stored in the HDFS file system first; then, the document is transformed into the high-dimensional matrix by the Singling algorithm of MapReduce. Minimum signature matrix is obtained by using the NewMinhash algorithm for the matrix, and the similarity degree is obtained approximately by using the Jaccard coefficient. A similar document was detected.

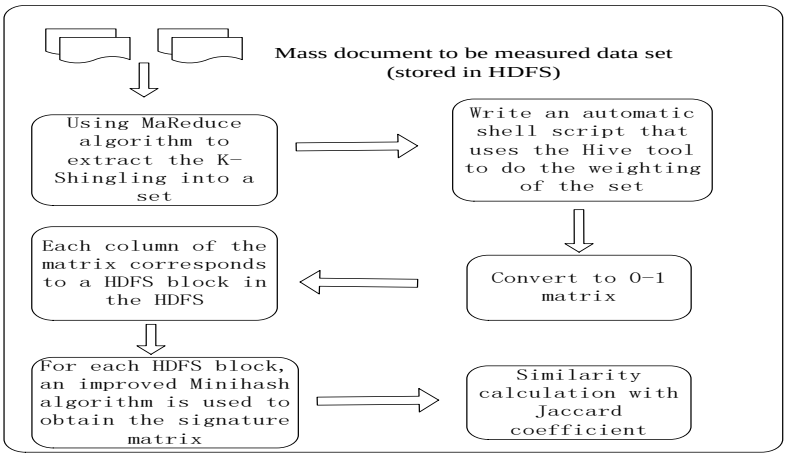

Figure 5. Detection process model

4.2. MapReduce-based Shingling algorithm

The traditional Shingling algorithm converts each document into a string as input, selects k value, intercepts the string to a substring of arbitrary k length, and transforms the document into a k-shingle collection of one or more times in the document. In general, the traditional small document $k=5$, large document $k=9$. However, some problems have been ignored. Firstly, there may be many meaningless function words in each text, which will become the result of the experimental noise. Secondly, the traditional Shingling algorithm cannot be applied in distributed environment. Therefore, we need to delete meaningless function words, then MapReduce to adapt to the distributed environment. For example, the string converted by document D is abcbe, where e is a meaningless function word. In the traditional algorithm, the selection of k should be large enough to ensure that arbitrary shingle is relatively low in any document, because the string of document D is relatively short. So if you choose $k=2$, then the set of all 2-shingle combinations in document D is {ab,bc,cb,be}, and after the function word is removed, the result is {ab,bc,cb}. The algorithm describes the steps as follows:

(1) Remove meaningless function words such as interrogative words, conjunctions, interjections, and then store them in HDFS.

(2) Design of Map function.

The input key and value are used as row identifiers and each line of content in the document. The output of the key is 1 to facilitate the Reduce function to handle the original output of value, and all need to be serialized. This way, the shuffle procedure in MapReduce automatically sorts and merges the value with key equals 1 so that each substring can be obtained, and the Reduce procedure can be easily merged into a document string.

(3) Design of Reduce function

Input key and value are the output of the Map function, load each line of the document, merge into a document string, and then generate 5-shingle according to k=5. Finally, each shingle is outputted to the HDFS so that the k-shingle of each document is obtained, and a HDFS block is taken in the Hadoop.

4.3. NewMinhash algorithm

Documents processed by the Shingling algorithm generate very high dimensional collections. For example, a 7.8K document is processed to generate a set of 7984 dimensions, the size of which is 44.5K. If there a large number of documents, there may be some problems such as slow calculation, memory overflow and so on. We use NewMinhash to reduce the dimension, and use the MapReduce framework to calculate the minimum hash signature. First, the feature matrix of the document must be stored according to the block, and each block corresponds to a column in the hash matrix. For example, the column vector of document S1 (1, 0, 0, 0, 1) and the column vector of document S2 (1, 0, 1, 1,) are stored in two files of Hadoop, respectively. It is easy to use the parallelization of Hadoop. Each Map task receives some columns and all hash functions, then calculates the minimum hash signatures of these given columns, and obtains the Jaccard similarity for the resulting minimum signature matrix. It is worth noting that when the minimum number of Hashi functions is relatively small, the signature matrix of the set matrix is very small after the compression of the set matrix. The traditional algorithm only pays attention to the element of the corresponding column equality, regards the equality as the similarity component, but neglects some very close elements, for these similar elements when seeking similarity, they also think they are similar. This way, the matrix similarity will be more accurate. Magnify the matrix and enlarge the element

with the absolute value of the difference of the difference of the difference of the difference of 1 to the larger one. The improved NewMinhash algorithm is as follows:

(1) First, we define the cluster of minimum hash function h1,h2,…,hn.

(2) Initialize ∞ for minimum signature matrix,SIG(i,c).

(3) Do the following for the block elements taken out of the distributed system:

if (element=0)

do nothing

else (element=1)

when i=1，2，…，n，using the minimum value between the original SIG(i,c) and hi(r) to replace the SIG(i,c).

The two columns in the minimum signature matrix that need to be compared are magnified. If the absolute value of the difference of the corresponding element is equal to 1, the small element is enlarged to a large element, but the original column is invariant.

4.4. Experiment and its results

In this part, we conduct the experiment based on the data set, SogouCS in 2008 that includes 11165 Chinese documents, 12 document categories, military, financial, and geographical areas, as shown in Table 3. Experiment 1 mainly verifies the validity of NewMinhash algorithm, which is the core part of the detection model. Experiment 2 verifies the accuracy of the proposed model compared with the traditional model. Experiment 3 verifies the extensibility of the model in distributed system.

Table 3. SogouCS Data set summary

Category	Military	Finance	Region	Property	Education	Technology
Document number	138	2179	2604	206	465	1038
Category	Car	Sports	Art	Entertainment	Game	Encyclopedia
Document number	344	718	883	1338	271	981

Experiment 1: Verify the effectiveness of NewMinhash algorithm.

NewMinhash is the core algorithm of the detection model proposed in this paper. It needs to be compared with the traditional algorithm Minhash to verify its validity. Because the data set is too large, the content is highly correlated, the verification process takes too long, and the validation of all the data sets has little effect on the validity of the algorithm. Therefore, the selected part of the data set can detect the effectiveness of the NewMinhash algorithm compared with the traditional algorithm; so, six documents from the entertainment type can be selected for verification. The experimental steps are as follows:

(1) Shingling algorithm for the six document S1, S2, S3, S4, S5, S6 extract set into 0-1 matrices. The actual values of the Jaccard similarity between any two documents in the matrix before the application of the algorithm are obtained respectively.

(2) In the distributed environment, after storing the matrix in the HDFS block, the Jaccard similarity is obtained by using the distributed NewMinhash algorithm for each HDFS block.

(3) The traditional Minhash algorithm is used to calculate the Jaccard similarity between documents, and compare with the similarity of the first two steps.

The results are shown in Table 4 and Figure 6. Table 4 shows the comparison between the two algorithm similarity values and the actual values between the six documents. It can be seen that the improved NewMinihash accuracy is closer to the actual value. For example, in ordinal number 5, the similarity of the Minhash algorithm of document S1 and S4 is 0.75 and 0.73, and the actual value of 0.74% New Minhash is more accurate. It is more efficient to select the similarity of document from the dataset. Figure 6 is the similarity curve between the two algorithms and the actual value. It can also be seen that the similarity curve of the NewMinhash algorithm is closer to the curve of the actual value.

Experiment 2: Detect the accuracy of the model.

Experiment 1 has verified the effectiveness of the core algorithm, and then the model is validated against the whole data set. The experimental steps are as follows:

Table 4. Two algorithms similarity

Serial number	Sim(Si，Sj)	Minhash	NewMinhash	Actual value
1	S1 S2	0.61	0.93	0.92
2	S1 S3	0.45	0.72	0.91
3	S1 S4	0.36	0.52	0.21
4	S1 S5	0.72	0.72	0.80
5	S1 S6	0.65	0.73	0.74
6	S2 S3	0.65	0.78	0.82
7	S2 S4	0.31	0.59	0.41
8	S2 S5	0.52	0.67	0.63
9	S2 S6	0.65	0.71	0.76
10	S3 S4	0.31	0.52	0.21
11	S3 S5	0.75	0.73	0.75
12	S3 S6	0.61	0.51	0.76
13	S4 S5	0.52	0.55	0.61
14	S4 S6	0.62	0.68	0.65
15	S5 S6	0.43	0.59	0.54

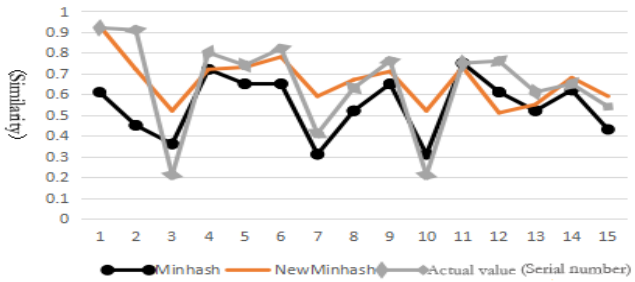

Figure 6. The similarity curves of the two algorithms and the actual values

(1) Because Chinese texts often contain some meaningless function words, in order to improve the accuracy of the experiment, the data set is preprocessed. Remove meaningless function words in the text. The reference to removing function words is 1208 function words in the Chinese disabled word list.

(2) According to the model proposed in this paper, when the 0-1 matrix is calculated, the similarity between any two documents is calculated, and the threshold is set, and the number of actual similarity pairs of documents is calculated.

(3) Then the final signature matrix is worked out, and the Jaccard similarity between each text is calculated separately. If the similarity exceeds the set threshold to indicate the similarity between the two documents, the number of similarity pairs for each category in the dataset is calculated.

The similarity model of traditional algorithm is used to calculate the number of similarity pairs, and the text similarity pairs are compared with the improved model. The similarity pairs larger than the threshold are selected, and the accuracy is obtained. The accuracy is the model similar logarithm / actual similar logarithm; its size reflects the accuracy of the model.

The results of the analysis of 12 document categories are shown in Table 5, which shows that the number of traditional similarity columns is less than that of the improved model, and the accuracy is smaller than that of the improved model. For instance, the number of traditional similarity pairs is less than that of the improved model, and the accuracy of the improved model is smaller than that of the improved model. It is proved that the accuracy of the document similarity detection model proposed in this paper is higher than that of the traditional model, and the similarity between text pairs can be calculated more accurately to select similar document pairs.

Table 5. Comparison of two models

Serial number	Document categories	The number of traditional similarities	The number of improved model similarities	The accuracy of the traditional model	The accuracy of the improved model
1	Military	720	765	0.79	0.84
2	Finance	28114	27070	0.77	0.83
3	Territories	48202	51214	0.80	0.85
4	House property	3454	4605	0.54	0.72
5	education	6824	6827	0.79	0.79
6	Science and technology	36941	44813	0.61	0.74
7	Car	3547	3697	0.70	0.70
8	sports	7124	7651	0.81	0.87
9	Literature and art	8412	8731	0.79	0.82
10	Amusement	15758	18599	0.61	0.72
11	Game	1694	1735	0.81	0.83
12	Encyclopedia	12148	12459	0.78	0.80

Experiment 3: Extending the test model

After the veracity of the model is verified, the expansibility is then verified. When using the traditional model to analyze the data set, it is found that it takes too long. If the amount of data is further increased, there will be insufficient memory and even a crash, so it is necessary to verify the expansibility of the improved model. With the increase of computer nodes, if the time is getting smaller, the scalability can be proved to be better. The number of computing nodes in the experimental platform is set to 2，4，6，8，10, respectively, and the nodes are configured the same host. Then, the data set is processed in experiment 2, and the result is shown in Figure 7. It can be seen that the processing time is greatly reduced with the increase of node data. It is proved that the model has good expansibility on Hadoop platform. When the amount of data is very large, the processing ability can be extended by increasing the number of nodes.

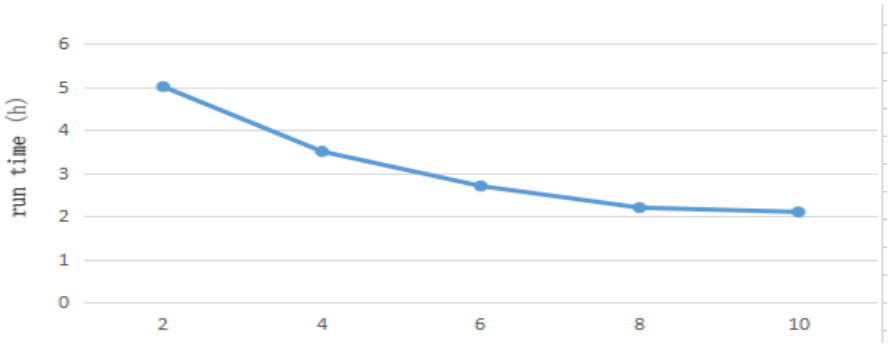

Figure 7. The run time of different nodes

5. **Conclusions**

In this paper, we design a high reliability Web log data mining scheme and put forward a kind of text similarity simulation detection model based on Hadoop. We use the tools such as Zookeeper, Sqoop and Hive in Hadoop ecosystem to combine traditional database MySQL. We propose a solution for mining user behavior log data, and design a highly reliable data mining scheme for network logs. Aiming at the existing user behavior network log data set, the filtering algorithm based on MapReduce is realized in the data preprocessing stage, the high-quality target data is cleaned, the data is analyzed by tools Sqoop and Hive, the PV ranking is excavated, IP number (number of visitors), traffic top ten IP and many other hidden information, provided to the enterprise decision-making. Then, compared with the traditional text similarity mining model, this paper proposes a model of text similarity based on Hadoop. The MapReduce-based Shingling algorithm and the distributed NewMinhash algorithm are designed to improve the accuracy of the similarity detection model. The simulation results of SogouCS news data are carried out to verify the effectiveness of NewMinhash under distributed, and the model accuracy and expansibility are verified. It proves that the model can select the similar documents in the document more accurately than the traditional model. It also has a good scalability in the distributed environment, as well as provides a certain reference value for solving the similarity detection of large-scale data in distributed environment.

Acknowledgements

This work was supported by the national natural science foundation of China ([2018] 61741124), central special funds to guide the development of local science and technology project ([2016] 4008), Project of Jiangsu Provincial Policy Guidance Program (Enterprise-University-Research Institute Cooperation)-Perspective Joint Research Program under Grant No.BY2016065-52,Taicang Key Project of Research and Development Plan(Industry Perspective and Common Key Technologies) under Grant No.TC2016GY07. What's more, we thank the anonymous reviewers sincerely for their significant and valuable feedback.

References

1. E.J. Chen, E. B. Jiang. "Review of Studies on Text Similarity Measures," [J]. Data Analysis and Knowledge Discovery,201 7,1(06):1-11.
2. F. J. Feng, J. P. Yao, X. S. Li, J. C. Ma. "Research on the Data Cleaning Framework," [J]. COMPUTER ENGINEERING&SOFTWARE,2017,38(12):193-196.
3. Gartner IT Glossary. "Dark Data," [EB/OL]. [2015-03-16]. http: / / www. gartner . com / it - glossary / dark - data.
4. G. M. Hu, L. Zhou, L. X. Ke. "Research on Hadoop-base Network Log Analysis System," [J]. Computer Knowledge and Technology,2010,6(22):6163-6164+6185.

5. D. X. Li. “Web Log Analysis Based on Data Mining,” [J]. Computer Knowledge and Technology,2011,7(25):6074-6075+6078.
6. Kumar N. “Approximate String Matching Algorithm,” [J]. International Journal on Computer Science and Engineering,2010，2（3）：641-644.
7. K. L. Shen, B. Shao, J. Du. “The Realization of Digital Resource Monitoring System Based on Network Log Analysis,” [J]. RESEARCH ON LIBRARY SCIENCE,2015(16):21-25.
8. S. M. Xie. “Forum Log Analysis Based on The Big Data Processing Technology Hadoop,” [D]. Jiangxi Agricultural University,2014.
9. Z. M. Xia, X. Liu. “A Similarity Algorithm for Chinese Text Based on Semantics,” [J]. JI SUAN JI YU XIAN DAI HUA,2015(04):6-9.
10. X. J. Xiang, Y. Gao, L. Shang, Y. B. Yang. “Parallel Text Categorization of Massive Text Based on Hadoop,” [J]. Computer Science,2011,38(10):184-188.
11. D. H. Yang, N. N. Li, H. Z. Wang, J. Z. Li, H. Gao. “The Optimization of the Big Data Cleaning Based on Task Merging,” [J]. CHINESE JOURNAL OF COMPUTER,2016,39(01):97-108.
12. F. Y. Yang, H. C. Liu. “Research on Hadoop Base Online Network Log Analysis System,” [J]. Computer Application and Software,2014,31(08):311-316.
13. Q. L. Yang. “Internet User Behavior Analysis Based on Web Log,” [D]. Huazhong University of Science &Technology,2013.
14. L. L. Zhang. “Research and Implementation of Chinese Text Categorization Based on Hadoop and SVM Algorithms,” [D]. Kunming University of Science and Technology,2015.
15. P. Z. Zou. “Website Evaluation Index and Construction Status Analysis,” [J]. Computer CD Software and Application,2012,20:151-155.

Yun Wu received the Ph.D. degree from Guizhou University, Guizhou, China, in 2009. Now he is an associate professor, graduated supervisor, and the member of China Computer Society. His research interests include Distributed Computing, Game Theory, Recommender System, Big Data and its Application.

Xin Ma is a master student in the College of Computer Science and Technology, Guizhou University. Her current research interests include Distributed System, and Recommender System

Guangqian Kong received the Ph.D. degree from Guizhou University, Guizhou, China, in 2009. Now he is an associate professor, graduate supervisor, and the member of China Computer Society. His research interests include Computer Network, Big Data and its Application

Bin Wang received his Ph.D. from the School of Automotive Studies in Tongji University, China. Now he is a professor of School of Mechanical Engineering at Yancheng Institute of Technology, China and a visiting scholar of School of Engine ering at Penn State Behrend. His current research interests include Automotive Drive/Transmission System, Electric Drive S ystem, High Performance Transmission, Power Loss of Transmission, Manufacturing based on the Internet of Things, Netw orked Manufacturing.

Xinwei Niu received his Ph.D. from the Department of Electrical and Computer Engineering in Florida International University, USA. Now he is a visiting assistant professor of Electrical and Computer Engineering at Penn State Behrend. His current research interests include High Performance Computing, Hardware Acceleration, Reconfigurable computing, Hardware Security, and Power-/thermal-aware computing.