Neural Comput & Applic (2017) 28 (Suppl 1):S327–S335
DOI 10.1007/s00521-016-2349-x

ORIGINAL ARTICLE

Research of network data mining based on reliability source under big data environment

Jinhai Li[1,2] · Youshi He[2] · Yunlei Ma[3]

Received: 9 March 2015 / Accepted: 9 May 2016 / Published online: 23 May 2016

Abstract In the era of big data, facing vast amounts of network data, only identifying the reliable data source can the researchers extract the original data that can be used in scientific research. Building reliable network data mining model based on the improvement of PageRank algorithm with applying each improved algorithm. Then the model is divided into three modules: the first, use PageRank and TrustRank to eliminate cheating webpages; then, refine webpages which related to research topic highly by TC-PageRank which combined with the topic relevancy between webpages and weight of time difference; finally, determine the authoritative webpages of the original data source by the improved HITS which considered the influence of the similarity between webpage and research topic and the amplification of webpage links to the authoritative webpages. Meanwhile, the partitioning of matrix operation based on MapReduce reduces the time and space complexity of the algorithms. And the feasibility and accuracy of the method are verified by comparative analysis of the algorithms.

Keywords Big data · The original data · PageRank · TC-PageRank · HITS · MapReduce

✉ Jinhai Li
ljh-hk@163.com

[1] Taizhou University, Taizhou 225300, China

[2] School of Management, Jiangsu University, Zhenjiang 212013, China

[3] Faculty of Science, Jiangsu University, Zhenjiang 212013, China

1 Introduction

Data are the indispensable part of the scientific research which runs through the whole process of the research. It includes the original data which used in the experiment and the experimental data which obtained from experiment. For example, in the study of usefulness of online reviews, the original data are the data set of online reviews that collected from network, and the experimental data are the influence factors of the usefulness that got from experiment.

Currently, most researches of data mining focus on the analysis of the experimental data, including the statistical analysis of data and the interaction among the variables, or extract information from data set. For example, Malone et al. [1] extracted using rule from Kohonen self-organizing maps by data mining technique. Mohanty et al. [2] used data mining technique to classify and detect breast cancer from mammograms, and the technique also can be used in threat detection [3]. Small and Medsker [4] also summarized the technologies and applications of data mining. As the purpose of scientific research, the importance of experimental data is not doubt. However, as the basis of scientific research, the original data are not only an important part of scientific research but also may affect the experimental data. But the original data usually get from the existing data sets [5] or the web crawler [6]. For example, Ahuja et al. [7] thought web crawlers were full text search engines which assisted users in navigating the web and these web crawlers could also be used in further research activities. Xu et al. [8] proposed a user-oriented web crawler that adaptively acquired user-desired content on the Internet to meet the specific online data source acquisition needs of e-health researchers. However, the original data that gets from these ways has many defects. For example, the existing data sets were systemized by the

predecessors, although they can avoid the problem of the data acquisition, the timeliness is poorer. Only few research fields have existing data sets now, and some data sets are confidential. Not all researchers can get them easily. It is not conducive to the development of the scientific research. Moreover, the existing data sets contain less data and cannot random expand according to the demand of the experiment. But the size of the original data set affects the accuracy and comprehensiveness of the experimental data in the research. However, only through plenty of data can many experiments obtain a comprehensive experiment conclusion and reduce the error of the experiment.

It is especially reflected in the experimental analysis based on the network data, such as early warning of network public opinion [9], feature extraction of online reviews [10] and so on. Only overall collecting the network data can we truly reflect the level of network public opinion and the characteristics of the products involved in online reviews. However, everyone is equal in front of network data. Anyone can mine the data that needed through the web crawler. At the same time, as the development of some fields of scientific research, the existing data sets may not meet the needs of the present experiments. For example, in the early research of early warning of network public opinion, the data set major includes portal sites, famous BBS and so on. After Twitter was launched in 2006, it is becoming more and more popular in Internet users. The network popular words which were born in Twitter are also rapidly popular in network; Twitter effect is gradually formed. Twitter has become a rendezvous for comment of Internet users. So the early warning research of network public opinion needs the data from the Twitter. And the collection of network data can be added into existing data sets. But the data set which is got by this method contains a lot of noise data. It goes against the pretreatment of the data in the experiment and also affects the accuracy of experimental results. Because the network data belong to the user generated content of the era of social media under Web2.0 Environment, its reliability was questioned by the academia [11]. Therefore, only the reliable network data which were identified and mined can be used for scientific research.

2 The characteristics of network data under big data environment

The number of network data is so huge in the era of big data that the comprehensive collection of network data is a mission impossible. Mining reliable and authoritative network data is a viable way to acquire the original data. Network data as the main part of big data have the main characteristics of it [12], as shown in Fig. 1.

Recently, the network data volume has reached PB level and even ZB level. Baidu increases 10 TB of data every day, while the system needs to deal with 1 PB of data. Taobao produces >50 TB of the active data a day. Vast network data comes from e-commerce platform, web portal, network BBS, micro blogging, blog and so on. It includes the structured, semi-structured and unstructured data of text, audio and video, web logs. Its types are huge and variable and will increase with the generation of new technology. However, in the large number and wide kinds of network data, the really valuable data is extremely scarce. Network data also increase at an annual rate of 59 %, so if not handled in time, more and more data will be deposited.

In the era of big data, network data have penetrated into every industry and area and become important production factor gradually. It means more fields can use the data from the network for scientific research. The rigor of scientific research requires that not only the experimental data come from real experiment but also the original data for experiment are reliable. The existing data sets were censored and controlled strictly by corresponding researchers, so these data sets are reliable and important after processed. But the network data are almost without sifting and filtering. The public can express their views freely to achieve dissemination of information publicly. The characteristic of the openness of social media is reflected here. Though network data are all-inclusive, it has a large number of false and useless information due to lacking of strict control. Therefore, the reliability of the source of network data was a controversial topic. And in the face of vast network data, how to determine the important data for scientific research is also a question that needs to be considered.

3 Mining process of reliable network data

If we want to obtain the reliable network data from the vast network data, except excellent mining tool of network data, confirming the data source of collecting is the key. Different from traditional methods of web crawler, in the paper we build reliable network data mining model based

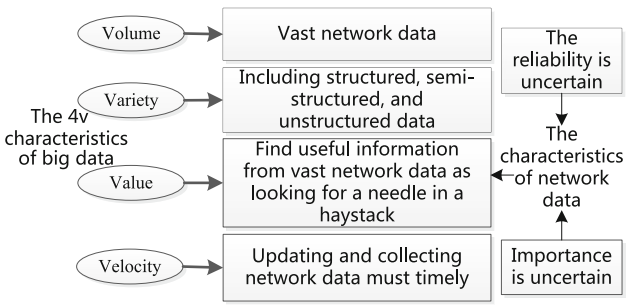

Fig. 1 Similarities between big data and network data

on the improvement of PageRank algorithm with applying each improved algorithm. Then the model is divided into three modules. As shown in Fig. 2, the 4, 5, 6 are the webpages which associated with the research field. First, identify the reliable webpages 3, 4, 5, 6; next, extract the webpages 4, 5, 6 which are associated with the research field from them; finally, determine the authoritative webpage 6 as the data source of the original data for scientific research from the webpages 4, 5, 6. And the module 1 uses the improved PageRank algorithm, the module 2 uses the improved TC-PageRank algorithm, and the module 3 uses the improved HITS algorithm.

3.1 The improved PageRank algorithm

PageRank algorithm is the key technology to success for Google search. It is a method of sorting the importance of webpage [13]. But cheaters change the importance of cheating webpages through cheating PageRank algorithm based on link cheating in the driven of economic interests, which makes traditional PageRank algorithm cannot get satisfied results. PageRank algorithm is simple and shown as follows:

$$p' = \beta Mp + (1-\beta)\mathrm{e}/n \tag{1}$$

The PageRank algorithm which used in the form of vector is good for calculation of PageRank of webpage with massive number of nodes in the network, where β $(0 < \beta < 1)$ is the damping coefficient, usually sets 0.85. M is a transfer matrix ($n \times n$ matrix used to describe the link between nodes). P is a vector which is composed of the value of PageRank of all the nodes in the current iteration. n is the number of all the nodes in the WEB. e is column vector of n dimension.

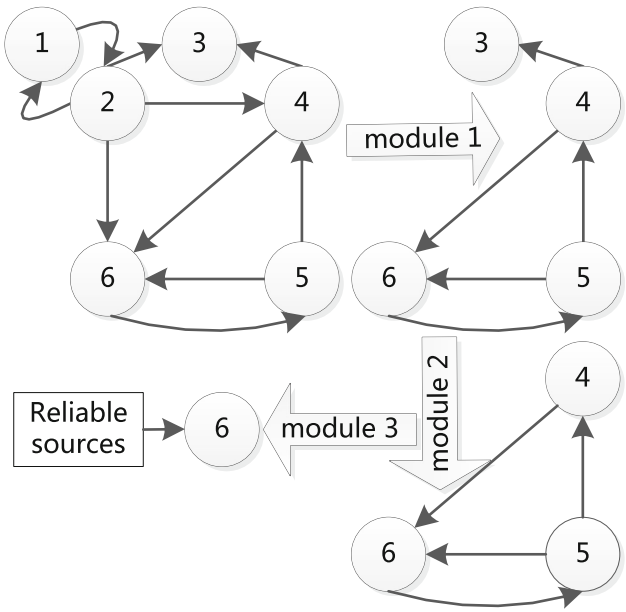

Fig. 2 Mining process of reliable data source of the original data for scientific research

In order to solve the link cheating and identify reliable webpage in the network, TrustRank algorithm is introduced to improve PageRank algorithm in module 1. TrustRank is deformation of subject-oriented PageRank. The "subject" is not a theme of web content but a reliable set of webpages. The reason that TrustRank can avoid link cheating is that cheating webpage can link to reliable webpage automatic, but not the opposite. So we need to choose a reliable set of webpages when set collection of random jump.

The process of the improved PageRank algorithm is: (1) Calculating the value of traditional PageRank p. (2) Calculating the value of TrustRank t. (3) Setting a threshold:

$$l_1 = (p-t)/p \begin{cases} l_1 > 0.5 & \text{cheating webpage} \\ l_1 \le 0.5 & \text{normal webpage} \end{cases} \tag{2}$$

The threshold l_1 sets closer to 0, and the punishment of link cheating is more rigorous. l_1 expresses the proportion of waste garbage (junk quality) of the value of PageRank of the webpage. In this way, the cheating webpages of high junk quality can be removed in the module 1.

3.2 The improved TC-PageRank algorithm

Seen from formula (1), because of the PageRank algorithm only uses the link structure of the web to sort, the algorithm exists faults of topic drift and laying particular stress on the old webpages [14]. Though the module 1 rules out cheating webpages in a great degree, the amount of webpages is still too large and contains large amounts of unrelated webpages. Meanwhile, the paper is about mining the reliable original data for scientific research. The rigor of scientific research needs the original data with real time. So we should eliminate obsolete data or reduce their influences.

TC-PageRank algorithm can solve the problem of topic drift effectively by assigning weight according to the relevance of webpage topic [15]. The relevance of webpage topic is calculated through the vector space model. Assume the document vector of the webpages u and v as $U = (u_1, u_2, \ldots, u_m)$, $V = (v_1, v_2, \ldots, v_m)$. u_i and v_i denote parameter values of feature word i in their webpages (calculated through the TF.IDF).

Then use $W(c)$ to denote the weight of webpage v among $F(u)$. $F(u)$ is representative of all outlinks of u.

$$w(v,u) = \frac{U \cdot V}{|U| \times |V|} = \frac{\sum_{i=1}^{m} u_i v_i}{\sqrt{\sum_{i=1}^{m} u_i^2 \sum_{i=1}^{m} v_i^2}} \tag{3}$$

$$W(c) = \frac{W(v,u)}{\sum_{p \in F(u)} W(p,u)} \tag{4}$$

In view of the idea of TC-PageRank algorithm improving topic drift problem, the paper also solves problem of laying particular stress on the old webpages through reducing the topic relevance of it by idea of reducing weight. Assume that the greater the time difference of t_1 (the time of webpage be searched) and t_2 (the time of webpage be updated latest), the lower the relevance of webpage topic. Introducing time difference to weighting function $W(t)$:

$$W(t) = \frac{d}{t_1 - t_2} \tag{5}$$

Obviously, time difference is inversely proportional to the weight. d is a constant based on actual research field. t_1 and t_2 are calculated by the day. TC-PageRank algorithm which introduced weighting function $W(t)$ of time difference is as follows:

$$p' = \beta Mp \times W(c) \times W(t) + (1 - \beta)\mathrm{e}/n \tag{6}$$

The paper aims at studying the original data for scientific research in the network. There is useless information such as advertisements in most websites. The keywords in the AD will affect the measure of topic relevance of webpages to a certain extent. In addition to introducing time weighting function into TC-PageRank algorithm, according to the characteristic of the HTML document structure, TF.IDF algorithm of feature works was adjusted. A webpage is divided into several pieces. Different weight value is given to the feature words in different HTML tags. The webpage was parsed into a DOM tree. Then according to HTML tags such as <meta keywords>, <meta description>, <title>, <p>, <h1>, the DOM tree was parsed into DOM subtrees. Every DOM subtree represents a single piece of information. We assign different weighting factor w_i to the feature word i with different HTML tags. So parameter value of i in webpage u is $u_i' = w_i u_i$.

The paper sets the weighting factor w_i of different HTML tags as:

$$w_i = \begin{cases} 2, & <\text{title}> \\ 1.8, & <h1> \\ 1.5, & <\text{meta}> \\ 1, & \text{others} \end{cases}$$

Setting the threshold l_2, the webpage will be judged to the highly relevant webpages with research topic when TC-PageRank values of webpage greater than or equal to l_2. In this way, the module 2 can screen out the highly relevant webpages with research topic.

3.3 The improved HITS algorithm

The difference between HITS algorithm and PageRank algorithm is that PageRank views webpages as the only one-dimensional importance but HITS views webpages as the two-dimensional importance. The role of module 3 is to extract a certain amount of authoritative webpages as the source of the original data for scientific research from a large number of webpages with related research topic which are established.

HITS is a mining algorithm of web structure. Find out the authorities and hubs in the web by analyzing the link relations between webpages. Authorities are the authoritative webpages of related research topic. And this is also the source of data that module 3 needs. Hubs do not provide information themself, but contain a number of webpages which link to the authoritative webpages [16]. Process of the HITS algorithm is: (1) Building adjacency graph of web. To select the first n as the root set R in the webpages of related research topics from module 2. Expanding R to base set S through adding the webpages that R point and point to R. Set the number of webpages in the base set S is s. If a webpage i points to j, then $L_{ij} = 1$; otherwise, $L_{ij} = 0$ in the link matrix L of $s \times s$. The similarity between the link matrix L and the transfer matrix M is that both are all about link relationship between the webpages. But L is 1–0 matrix and M is the score matrix. M_{ij} is from 1 divided by the degree of the webpage i. (2) Calculating the value of authorities and hubs by the formula.

We can see from the process of HITS algorithm that when building adjacency graph of web, the root set is related to the topic. The expansion of the base set only considers the links to the root set but ignores the topic relevance, which will introduce a large number of webpages of little relevance to the topic. The authoritative webpages that finally obtained are not highly correlated with the research. Based on the idea of TC-PageRank algorithm, we introduce the judgment of webpage topic in the process of expansion of base set. Let the webpages that be elected to base set be highly relevant to the topic.

Because the links between the new webpages and other webpages are less, the authorities and hubs of the new webpages are smaller. But for the real-time requirement of the original data for scientific research, the paper argues that the determination of authoritative webpages not only needs to consider quantity of links but also needs to consider growth situation of links. If the growth of link is very fast, it means that it has certain reference and attention by more scholars. Based on the above two improvements, the paper puts forward a HITS algorithm hybrid the similarity of webpages and the amplification of links.

In order to calculate conveniently, building adjacency graph of web in HITS algorithm should also be based on vector space model. We express research topic and web content in vector, research topic: $t = (t_1, t_2, \ldots, t_n)$. Process of the improved HITS algorithm is as follows: (1) selecting

the root set R, calculate the similarity between webpage p in the root set R and research topic t. If the similarity is less than the threshold l_3, filters webpage p, otherwise keeps p. (2) expanding the root set R. When adding webpage p to base set S, in addition to considering the relationship of links between webpages, we will calculate the similarity between the webpages which will join base set and research topics. If the similarity is less than the threshold l_3, filter webpage p; otherwise, join the base set S. (3) The value of authorities and hubs was calculated based on the consideration of the weight of the topic similarity and link amplification of webpage. The calculation of similarity is to express each webpage p as vector form: $p_i = \{(p_1^i, f_1^i), (p_2^i, f_2^i), \ldots, (p_n^i, f_n^i)\}$; p_j^i expresses the jth key of the ith webpage. f_j^i expresses its frequency in webpage. If a feature word p_j^i of webpage corresponds to research topics t_k, then $T_k = 1$, otherwise $T_k = 0$. This is because the number of feature words of research topics is far less than feature words of a webpage. Use vector cosine to calculate similarity between webpage and research topic.

$$s(p_i, t) = \frac{\sum_{k=1}^{n} \sum_{j=1}^{n} T_k \cdot f_j^i}{\sqrt{\sum_{k=1}^{n} T_k^2 \times \sum_{j=1}^{n} \left(f_j^i\right)^2}} \tag{7}$$

Here we use vector cosine to calculate similarity because of that the webpages which get from module two have certain relevance with the research topic already, and cosine vector is more suitable than TF.IDF.

The cycle of updating websites by search engine varies according to the weight of website. The websites that have great weight keep updating every day, but it takes a month to update all websites. It means that Google and other search engines will gather the whole webpages every other month. The webpages are re-ranked according to the formed topology. Through the following formulas, we use the changes of webpage link to calculate the amplification of webpage link.

$$l_p^{\text{in}} = \frac{l_t^{\text{in}} - l_{t-1}^{\text{in}}}{l_{t-1}^{\text{in}}} \tag{8}$$

$$l_p^{\text{out}} = \frac{l_t^{\text{out}} - l_{t-1}^{\text{out}}}{l_{t-1}^{\text{out}}} \tag{9}$$

l_p^{in} expresses the amplification of links which link to webpage p. l_p^{out} expresses the amplification of links which webpage p links to others. l_t^{in} expresses the number of links that link to webpage p in the tth collection cycle. l_t^{out} expresses the number of links that webpage p links to others in the tth collection cycle. And we use l_{in} to express the vector of l_p^{in} of all webpages and l_{out} to express the vector of l_p^{out} of all webpages. The formula of the improved HITS algorithm is:

$$a = \lambda L^T L a \times s + \mu l_{\text{in}} \tag{10}$$

$$h = \lambda L L^T h \times s + \mu l_{\text{out}} \tag{11}$$

λ and μ are the weight factors. They are used to balance the impact on the webpage between the topic relevance of webpage and amplification of webpage link. L^{T} is the transposed matrix of link matrix L. It means when there is link from webpage j to i, then $L_{ij}^{\text{T}} = 1$, otherwise $L_{ij}^{\text{T}} = 0$. Set the threshold l_4, and regard webpage as the authoritative webpage related to research topic when $a \geq l_4$. In this way, the module 3 can extract the authoritative webpage related to research topic.

4 The simulation experiment

The paper regards the Google search engine as the source of the experimental data and online reviews as the research topic. If we need to study the influencing factors of the usefulness of online reviews, take online reviews of mobile phone products for example; then, set the research topic t = (camera, screen, price, systems, brand, resolution,…, battery capacity). We use the above methods to determine reliable data source for research of the usefulness of online reviews about mobile phone products.

We used web crawler tool Heritrix to collect the webpages of research topic through using the model BroadScope on Google, and we set it as stop condition when collecting 200,000 pages. Then we performed a series of operations such as webpage preprocessing and text preprocessing, and then entered the required data into the calculation module. Finally, verify the accuracy of the model according to contrast analysis between the rankings of Google search and simulation results, as shown in Fig. 3.

4.1 The experimental data preprocessing

Experimental webpages are collected on the Google search engine based on the research topic. There is a lot of noise information including promotion and advertising of Google. Webpage preprocessing is used for cleaning these webpages. The number of webpages is 150,000 after cleaning. Webpage preprocessing also includes building parsing module of HTML document. It will parse a webpage to generate a DOM tree. The text preprocessing is used to establish feature vector of webpage. And through web logs of search engine we record the difference of the search time t_1 and latest updated time t_2 of webpage, and the amplification of webpage links between the two searches.

After webpage preprocessing, the next step is to count the forward and back links of webpage, and then establish

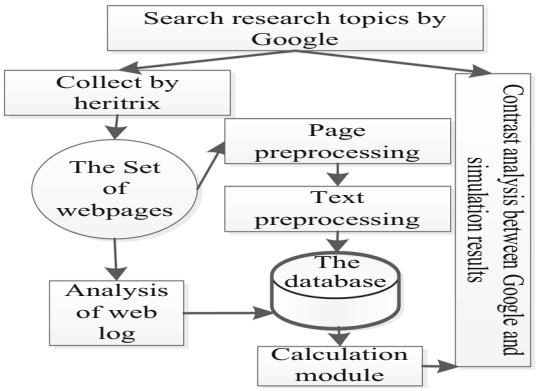

Fig. 3 Flowchart of experiment

transfer matrix M whose dimension is 150,000 × 150,000. Part of the transfer matrix:

$$M = \begin{vmatrix} 0 & 1/12 & 0 & \ldots & 0 & 0 \\ 0 & 0 & 1/10 & 0 & \ldots & 1/8 \\ 1/9 & 0 & \ldots & 1/13 & 0 & \ldots \\ 1/9 & \ldots & 0 & 0 & 1 & 0 \\ \ldots & 1/12 & 0 & \ldots & 0 & 0 \\ 0 & 0 & \ldots & 0 & 0 & 0 \end{vmatrix}$$

$M_{31} = 1/9$ expresses the first webpage has nine outlinks and links to the third webpage. Values on the diagonal of matrix are all 0. The transfer matrix is very sparse through observation. Each webpage only has 10 outlinks on average. In terms of 150,000 webpages in this experiment, the nonzero elements only accounts for 1/15,000 of matrix M.

All kinds of data are stored in the database in the form of Table 1 after preprocessing.

4.2 Matrix operation based on MapReduce

Matrix operation in the experiment chooses MapReduce [17], because the number of iterations of PageRank algorithm is too much, and the time and space complexity are large. Matrix operation based on MapReduce can get the performance boost of time and space through reducing the number of iterations by matrix block.

In formula (1), vector p represents all the value of PageRank. In this experiment, the dimension of the p is only 150,000. But the implementation of the system is under the environment of the whole web, the dimension of p is hundreds of millions at this moment, so it cannot be put into the memory directly. And transfer matrix M stored by columns is based on consideration of efficiency. Each column of M will be related to each component of p'. The component of p' is not stored in memory when we put one item into one component of p', so the algorithm needs to convert page into memory at the time of adding one item. It creates a memory thrashing that will increase the computation time with order of magnitude.

Based on this, we will divide transfer matrix M into k^2 blocks and vector p into k blocks. Block method is shown in Fig. 4 (other parts of the algorithm are not shown in the figure).

Set k^2 Map tasks on the basis of the number of blocks of M. Every Map task processes one block M_{ij} of the transfer matrix M and one block p_j of vector p (according to the rule of matrix–vector multiplication, j must be the same), where every block p_j of p inputs to k different Map tasks which deal with M_{ij} ($i = 1, 2\ldots k$). When dealing with M_{ij}, p_j and p_j' are kept in memory. All production of M_{ij} and p_j will only be used for calculation of p_j', so p will put into the algorithm for k times, and each block of M will input only once. The size of vector p relatives to the transfer matrix M is negligible, so it reduces the complexity of the algorithm greatly. The data size that Map task outputs to Reduce task is reduced, because the Map task repeats combination operation.

4.3 The steps and results of experimental simulation

We interconnect 20PC with Intel core i5-3.1 Ghz quad-core and memory of 4 G through 100 m/s switch to build a distributed environment. Set the difference of PageRank of a certain node less than or equal to 10^{-6} as the iteration convergence condition. Set the threshold of testing a webpage of cheating in the module 1 $l_1 = 0.5$. The initialization vector is $v = (1, 1,\ldots, 1)^{\mathrm{T}}$. It means the value of PageRank of each webpage is 1 at the beginning of the iteration. Set the threshold of judging the relation of research topic in module 2 $l_2 = 2$. Set the weight factors in module 3 $\lambda = 0.8$, $\mu = 0.2$ (assume that the topic relevance of webpage is more important than the link amplification of webpage). Set the threshold of building adjacency graph of web $l_3 = 0.1$ and the threshold of determining the authority pages $l_4 = 4$.

In order to verify the effectiveness of the improved method in the paper, compare analysis the algorithm accuracy between the traditional algorithm and improved algorithm, and analyze the execution time of matrix operation based on MapReduce.

To verify the accuracy of the improved algorithm by comparing the similarity of the authoritative webpages sorting of the improved algorithm and the traditional algorithm with Google search sorting, the operation process of the model is as follows: (1) Using the l_2^{out} and l_2^{in} in Table 1 to count the forward and backward link of every webpage to establish transfer matrix M. Use formulae (1) and (2) to remove the cheating webpages of high junk quality. Based on the value of PageRank, we get data set of reliable webpages from data set of collected {4,11,13,19,…,4257,…,62,547,…,149,981} (the digital in the set represents the ID of the webpage). (2) Use the VSM

Table 1 Information table of webpage

Data information	Explain	Example
ID	Unique identifier of a webpage	1,2,…,90,000
URL	The path of the webpages on the Internet	http://mobile.zol.com.cn/
Path	The path of the webpages in the database	E:/MySQL/MySQLServer5.5/data/index.html
t_1	The search time	2014-03-10
t_2	The recent update time	24 Jan 2014
l_1^{out}	The number of outlinks when the last time to search	7
l_2^{out}	The number of outlinks when this time to search	12
l_1^{in}	The number of inlinks when the last time to search	55
l_2^{in}	The number of inlinks when this time to search	131
Outlinks	The ID of webpage which current webpage links out	5,18,459,568,785,5486,7851,25,426,85,214
VSM	The feature vector of webpage	(<camera,4>, <fever,7>, <screen,5>, <price,10> …)
IDF	The number of webpages have feature	(<camera,13,384>, <fever,9451>, <screen,12,512>, <price,25,483>…)
DOM	DOM tree index of HTML	<meta name = "keywords" content = "phone, apple">

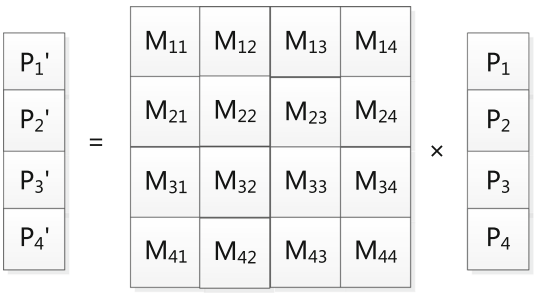

Fig. 4 Schematic diagram of matrix block

and IDF in Table 1 to calculate the correlation of topic w(u,v) between webpages by formula (3). Calculate the weight W(c) of webpage v in all the outlinks of u by formula (4). Use the t_1 and t_2 in Table 1 to calculate the weight $W(t)$ of the time difference of webpage update by formula (5). Combine the information of DOM tree in Table 1 and formula (6) to find the data set {4,11,19,…,4257,…,149,981} which related to research topic highly from the data set of reliable webpages. (3) Use the VSM in Table 1 to calculate similarity s between webpage and research topic by formula (7). Use the l_1^{out}, l_2^{out}, l_1^{in}, l_2^{in} in Table 1 to calculate amplification l_{out} and l_{in} of webpage p by formula (8) and (9). We calculated the value of authorities and hubs of data set of webpages by formula (10) and (11). Determine data set of authoritative webpages {4,19,…,4257,…} according to the value of authorities. Finally, find out the webpages from data set of authoritative webpages as a reliable source of original data for research through the three data of ID, URL, path in Table 1.

Similarity of sorting results of top 200 authoritative webpages is shown in Fig. 5.

Figure 5 shows that the accuracy of using the improved algorithm to calculate the authoritative webpages promotes from 20 % of the top 10 webpages to nearly 30 % gradually. The accuracy of the traditional algorithm model is unstable. The downtrend of accuracy of the traditional algorithm model is more apparent than the improved algorithm as the increase in the number of webpages. Seeing from the two downtrend line of accuracy, the convergence speed of accuracy of the improved algorithm is about twice as fast. It mainly thanks to consider the relevance of webpage topic, freshness of the webpage and the amplification of link impacting on webpages.

Three modules all include the matrix calculation, where the calculation of PageRank algorithm involves the most number of iterations in module 1. According to iterative convergence condition and different number of webpages,

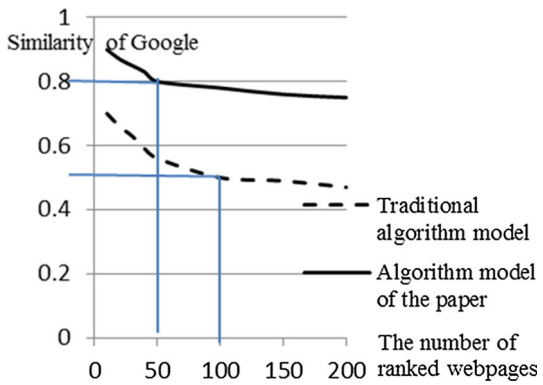

Fig. 5 Similarity of sorting results of authoritative webpages

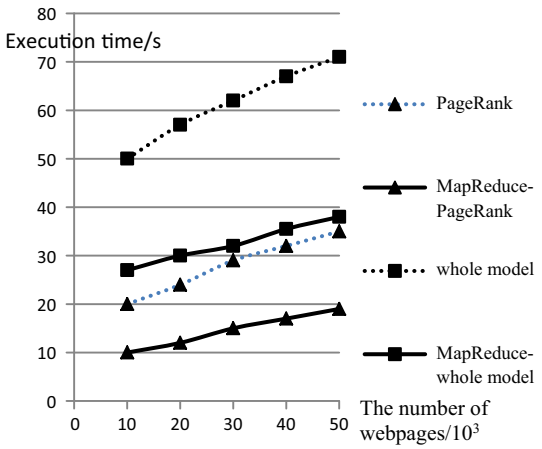

Fig. 6 Execution time under the different operation environment

we verify the execution time of the improve PageRank algorithm under the conditions of traditional operation and matrix operation based on MapReduce. And we also analyze the consumption time of whole model under the environment of traditional operation and matrix operation based on MapReduce. The results are shown in Fig. 6.

From Fig. 6, we can see that the execution time of matrix operation based on MapReduce is much smaller than the time under the environment of traditional computing whether operating a single task of PageRank or computing the whole model. And we also can find that the increase rate of the execution time is smaller based on MapReduce as the increase in the number of webpages. It shows that the matrix operation based on MapReduce has the advantage of low time complexity.

5 Conclusion and prospect

The paper starts from the actual demand that to get the original data for scientific research currently is difficult. The paper relied on the advantage of mass and wide coverage of network data to divide the acquisition of reliable original data for scientific research into three modules. We obtain reliable webpages through discriminating cheating webpages, then refine highly related webpages with research topic, finally determine authoritative webpages as the source of the original data for scientific research from the related webpages. The three modules run step by step. The model not only guarantees data set of webpages closely related to the research topic but also guarantees authority of data set of webpages.

The next step is to extract information from these data sets after obtaining reliable data sources for scientific research. These will be presented to the researchers in the form of more intuitive and normalization, which avoiding the problem of having a large number of data sources but do not know how to use.

But due to the lack of related research in academia at present, there is not a reliable criterion to test the accuracy of the experimental simulation. We can only verify the accuracy of the results through manual work. The data size is small in experiment. So we can finish inspection through manual work. But if we put the model in the system based on the whole web environment, it is impossible to verify the accuracy of the results through manual work. It is a key to improve the research with the help of a simple test method.

Acknowledgments This study was funded by the National Natural Science Foundation of China (71302087) and Graduate Innovative Projects of Jiangsu Province in 2014 (KYZZ_0287).

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Human and animal rights This article does not contain any studies with human participants or animals performed by any of the authors.

Informed consent Informed consent was obtained from all individual participants included in the study.

References

1. Malone J, McGarry K, Wermter S et al (2006) Data mining using rule extraction from Kohonen self-organising maps [J]. Neural Comput Appl 15(1):9–17
2. Mohanty AK, Senapati MR, Lenka SK (2013) An improved data mining technique for classification and detection of breast cancer from mammograms [J]. Neural Comput Appl 22(1):303–310
3. Bhardwaj AK, Singh M (2015) Data mining-based integrated network traffic visualization framework for threat detection [J]. Neural Comput Appl 26(1):117–130
4. Small SG, Medsker L (2014) Review of information extraction technologies and applications [J]. Neural Comput Appl 25(3):533–548
5. Cao XY, Zhang X, Liu L et al (2014) Research on internet public opinion heat based on the response level of emergencies [J]. Chin J Manag Sci 22(3):82–89
6. Yin GP (2012) What online reviews are more useful by consumers' thought? [J]. Manag World 12:115–124
7. Ahuja MS, Bal DJS, Varnica B (2014) Web Crawler: extracting the web data [J]. Int J Comput Trends Technol 13(3):132–137
8. Xu S, Yoon HJ, Tourassi G (2014) A user-oriented web crawler for selectively acquiring online content in e-health research [J]. Bioinformatics 30(1):104–114
9. Si XM, Liu Y (2011) Influence of internet chat rooms on network public opinion [J]. J Internet Technol 12(3):393–398
10. Chen L, Qi L, Wang F (2012) Comparison of feature-level learning methods for mining online consumer reviews [J]. Expert Syst Appl 39(10):9588–9601
11. Stvilia B, Gasser L, Twidale MB et al (2007) A framework for information quality assessment [J]. J Am Soc Inform Sci Technol 58(12):1720–1733

12. Hilbert M, Lopez P (2011) The world's technological capacity to store, communicate, and compute information [J]. Science 332(6025):60–65
13. Page L, Brin S, Motwani R et al. (1998) The PageRank citation ranking: bringing order to the web [EB/OL]. http://ilpubs.Stanford.edu: 8090/422. Accessed 19 Dec 1998
14. Richardson M, Domingos P (2002) The intelligent surfer: probabilistic combination of link and content information in PageRank [J]. Adv Neural Inf Process Syst 14:673–680
15. Haveliwala TH (2002) Topic-sensitive PageRank [C]. In: Proceedings of the 11th international world wide web conference, Hawaii, pp 517–526
16. Chang Q, Zhou MQ, Geng GH (2007) PageRank and HITS-based web search [J]. Comput Technol Dev 18(7):77–79
17. Dean J, Ghemawat S (2004) MapReduce: simplified data processing on large clusters [C]. In: Proceedings of the 6th conference on symposium on operating systems design and implementation, USENIX Association