Acta Technica *62* (2017), No. 3B, 857–868

A study on the parallelization of pagerank algorithm based on big data

Jianrui Xu[1]

Abstract. In recent years, with the rapid development of science and technology, the application of network has a very wide range, beyond the traditional small data sets. It gets to big data with large scale, dimension, and type. There are some difficulties about the value of database for us. The execution efficiency of data mining and algorithm parallelization of data mining are facing a big challenge. At the same time, the platform usability of data mining algorithm is also facing the big challenge. In order to realize the high efficient and high quality operation based on the big data background and conclude the complete data mining system, the paper researches the PageRank algorithm, and builds the initial model and then gets the results. However, considering the influence of data quality factors from the uncertainties of the environment, the paper introduces an escape factor, and establishes the improved model. Taking the enterprise image data as an example, using robustness analysis to compare the initial model and improved model, at the last, the paper gets the result: in the complex various environment, the improved model has much more advantages. This research brings theory basement for the application of big data in real life, and lay the foundation of the development of the related areas.

Key words. PageRank, initial model, escape factor, improved model, robustness analysis.

1. Introduction

With the development of information technology, facing the data of TB even PB, search engine has become the hotpot in the field of data mining application. The difficulty of search engine is ordering mechanism, so it needs a ordering mechanism to link to the page link as support degree order, it's PageRank algorithm in this paper. If the support degree is higher, it will place at the front of the search results, and it is convenience to search a large number of information[1]. But when using PageRank algorithm to rank, the calculation efficiency is not only low, but also the process need to consume large amounts of CPU and computation resources. So exploring a ordering mechanism with high efficient[2].

In the PageRank algorithm, the web connection structure is considered as an

[1]Workshop 1 - Zhenjiang Vocational Technical College (Jiangsu Union Technical Institute Zhenjiang Branch), Zhenjiang Jiangsu, China, 212016; e-mail: zjxjr@126.com

important reason of ranking, so the reliance of each connection is the precondition of normal running of algorithm[3]. At the same time, this is a disadvantage of algorithm. Because not every connection can objectively reflect the content authenticity and value, the connection among web pages has artificial operability, and in order to get higher rankings, some people will use pointless waste connection to accumulate more links, so the connection with strong subjectivity can't reflect the rankings[4]. The paper research the parallelization of PageRank algorithm from the source of connection, aiming at these difficulties, it not only can improve the calculation efficiency but also can reduce the consumption of resources about algorithm, such as CPU.

2. The main algorithm of data mining in big data

Compared to big data, data mining is not a new technology. Data mining is actually a process of information extraction and knowledge discovery from a large number of random, fuzzy, noisy data sets. With the progress of computer science, data mining algorithms have been fully developed and applied. In the case of single small data, there are a lot of very good data mining software, and these software provides a wealth of data mining algorithms. In the big data background, some big data processing frameworks can be paralleled to implement some data mining algorithms on the big data set.

From the big classification, data mining algorithms can be divided into the following categories:

(1)Association rule analysis algorithm.

The aim of association analysis algorithm is to find frequent item sets, such as the examples of famous beer and diapers. Through the supermarket shopping data, people find that men usually buy beer at the same time also buy diapers. Thus narrowing the distance between the two kinds of goods can further improve sales. The association rules algorithm is represented by the Apriori algorithm and the FP-Growth algorithm. By setting the minimum support degree, the frequent item set is sought.

(2)Clustering algorithm

The purpose of clustering algorithm is to find out the similar data item set. The clustering data items have the similar properties. Clustering algorithm is represented by Kmeans algorithm and nearest neighbor algorithm (KNN) as the representative. For example, in the recommendation algorithm, the need for similar users to recommend similar products, here you can use the KNN algorithm to find similar users.

(3)Prediction and regression algorithm

The main purpose of the prediction and regression algorithm is to predict the trend of the data and the trend, it is represented by the linear regression algorithm and naive Bayesian algorithm. Through the training set to obtain the forecast model, according to the prediction model, the return value of a corresponding input value can be got. Linear regression algorithm, generally through the gradient descent method, finds the best matching model parameters.

(4)Index sorting algorithm is represented by Google PageRank algorithm. PageR-

ank algorithm is Google's patent, Google uses the algorithm to order the web search. The degree of support is the number of linking to a website. If the Degree of support is higher, the value of the web page is higher, and the order is higher. The specific algorithm is that the web is taken as the node of a graph, the link is taken as directional edge of a graph. The graph is expressed by adjacency matrix. Using the multiplication of a matrix and transposition, the result can be got.

3. PageRank algorithm

In the 7^{th} international conference, Page, proposed PageRank algorithm in April 1998. It is a offline web page ordering algorithm, and is a method that connect backward links and forward links to mark the web page[5]. The essence of PageRank algorithm is a probabilistic problem. How can judge the web page is high quality? It is that the web page has been pointed out by many websites. If there are many such websites, this web page is probably with high quality[6].

3.1. The establishment of the model

In the PageRank algorithm, a digraph $H = (V, E)$ represents the link relations among the web pages, a node in the graph represents a web page, so a set V is consist of all nodes in the graph $v1, v2, \cdots, vn$. E, a set of directed edge, represents a hyperlink.

The main idea of PageRank algorithm is: first of all, each node is given an initial weight; the value of PageRank represents it. $P(j)$ represents the value of PageRank of nodevi. If there is a directed edge from node vi to nodevj, this directed edge points vi tovj, it means that node vi vote for nodevj. If $D(i)$ represents the number of directed edge from node i, so the PageRank value of $P(i)/D(i)$ represents the contribution degree. The calculation process of PageRank is following:

Each node vj is given an initial PageRank value.

$P_k(i) = (1-d) + d \times \sum_{(Vj,Vi)\in E} \frac{P_{k-1}(j)}{C(j)}; 0 < i \leq n$ (1)

After k iteration, the PageRank value of each node is

$$P_k(i) = (1-d) + d \times \sum_{(Vj,Vi)\in E} \frac{P_{k-1}(j)}{C(j)}; 0 < i \leq n \tag{1}$$

According to the above iteration manner to several rounds of iteration, if the PageRank value is convergence, or the sum of all node PageRank values are less than self-defined threshold, so the process can be stopped the iteration, at the same time the web pages are ordered according to the final PageRank value of each node. d represents the probability that the user link into a web page, it is named damping factor$(0 < d < 1)$.

3.2. The improvement of the model

For the initial model, the following summation formula represents the support degree of the PageRank page.

$$M(i) \tag{2}$$

$M(i)$ is PageRank of i, $D(i)$ is all web pages that are linked to i.

However, the formula ignores the quality of web sites, all sites are equal, in fact, there is different among the links. If the rank is only relied on the number of connections, it will be error inevitably. If a site is quoted by several famous links, the rank will be higher than other links that are quoted by unknown websites. It matches what people want. It needs to improve the formula to distinguish the links; the most common method is that each link is given a weight value. For example, the page of directed page I, the importance of each page is$M(j)/N$, there are N links pointing to page E, so the following formula is got.

(4)The constant C is for standardizing the results. From the formula, we can get that if the PageRank value of one web page is higher than other web pages, it means that that web page is quoted by many important websites or by large number of unknown websites. These two cases are in line with the expectations.

4. Solution of the model

In the second section, the specific process of PageRank algorithm can be represented with a directed graph. The directed graph containing coordinates can also be represented with adjacency matrix. If a page can link to another page, so it is expressed with chmetcnv Unit Namein Source Value Has Space True Negative False Number Type1TCSC01 in the matrix, if not, it is expressed with 0. For example, web references are shown by the following matrix.

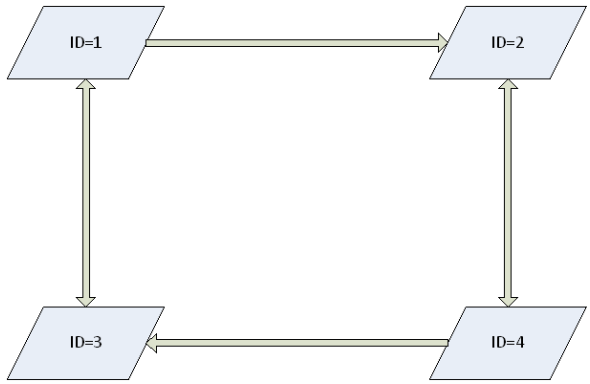

Fig. 1. Matrix graph

If page I link to page E, 1 expresses link. If not, 0 expresses it. So in the matrix, $n[i][j] = 1$, otherwise, in the matrix, $n[i][j] = 0$, adjacent matrix can be expressed:

$$N = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 \end{pmatrix}$$

In the matrix N, the number means if the node of value 1 is in the line I, it is the website linked by website I. If the value is 0, it means that no websites can be linked. If the node is in the line J, it means that it is the website linked by website J, if not; it means no websites can be linked.

After matrix changing, the matrix M can be got, at the same time then the transposed matrix M^T can be got.

$$M^T = \begin{pmatrix} \frac{1}{2} & 0 & \frac{1}{2} & \frac{1}{3} \\ 0 & 1 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & \frac{1}{3} \\ 0 & 0 & \frac{1}{2} & \frac{1}{3} \end{pmatrix}$$

In the matrixM^T, each value indicates the probability that network user link different web pages, and the initial probability is same. In this example, the probability is quarter. The column vectorU_0, which is n dimension and the value is$1/n$, is right multiplied by the transposed matrixM^T, then the transfer probability matrix of browsing the web can be got:

$$U1 = M^T U0 = \begin{pmatrix} \frac{1}{2} & 0 & \frac{1}{2} & \frac{1}{3} \\ 0 & 1 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & \frac{1}{3} \\ 0 & 0 & \frac{1}{2} & \frac{1}{3} \end{pmatrix} \begin{pmatrix} \frac{1}{4} \\ \frac{1}{4} \\ \frac{1}{4} \\ \frac{1}{4} \end{pmatrix} = \begin{pmatrix} \frac{1}{3} \\ \frac{1}{4} \\ \frac{5}{24} \\ \frac{5}{24} \end{pmatrix}$$

Then the transfer probability matrix $U1$ is right multiplied by$N$, the $U2$ is got. The loop iteration can be stopped until U is convergence.

Un (5)

After the operation, the final value in of the column vector is the PageRank value of corresponding web page. There are also some limitations in this process. One problem is that the precondition of the former process is that the figure must be strongly connected. It can also say that internet users can arbitrarily browse the web without limitation. Otherwise if browsing is ended, the value in the Un is zero. So the whole process will become meaningless. The other problem is that a node in the network (such as node b in the following Figure) may be not link to other web pages, and only has links to own . As shown below:

The Figure simply expresses users can't jump to other nodes when internet users jump to node b. The probability of transferring to node b is 1, at the same time the probability of transferring to other nodes is 0. This kind of situation is called Rank Sink.

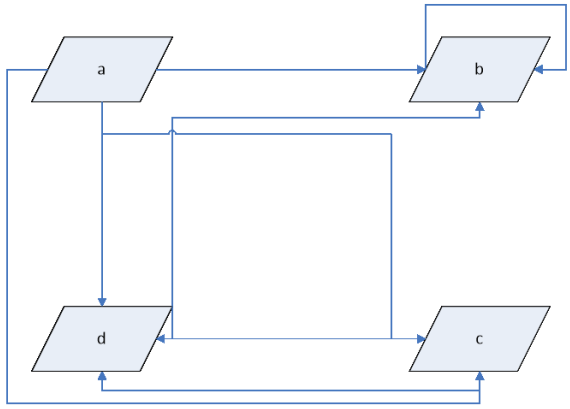

Fig. 2. Network nodes connection

5. The optimization of the model

5.1. The introduction of escape factor

In this section, two problems in the third section will be solved, at the same time the model will be optimized. First of all, for the probability browsed by Internet users, users may be not jump among several websites or provided connection but choose another path or directly input website to jump, so the concept of escape factor is introduced. Users escape several provided given websites. It means users' browsing can not be expressed by the node of original directed graph. Assume that if the user provides links to web sites continue to jump, the probability is p, so the probability of escape is 1 - p. The probability of transfer can be set and adjusted artificially according to the statistical results. Therefore, the original iterative formula can be transformed into:

$$PR\,(T) \tag{3}$$

In addition, if the adjacent matrix is a sparse matrix, it will cause a lot of memory and CPU resources waste with the above method. Hence, direct matrix multiplication is not optimal; the method can be replaced by another equivalent and incremental iteration calculation.

Through above analysis, the main operation of PageRank algorithm is matrix multiplication. But, the computation of matrix multiplication has high complexity. In addition, figure composed of web pages may be sparse graph. In other words, adjacency matrix is sparse matrix. If sparse matrix calculation is as common as matrix, the matrix multiplication can be directly executed.

Therefore, the calculation of transition probability matrix is not used, and another equivalent computation.

$$PR\,(T) \tag{4}$$

$PR(T)$ is PageRank value of web page T, tirepresents that web page T is linked by web page i. C (ti) represents the link number of web page i. a is the escape factor.

The probability of equation is optimized by this method. But in this way it will waste a lot of memory space. This is because sparse matrix is different from common matrix; there is lots of invalid data in the sparse matrix. So it needs right method to store web pages and to point to the links of web pages. The form is< URL, out link >, key is the current web page, valueis a list of web links.

5.2. PageRank parallelization process analysis

When the PageRank algorithm is in the parallelization, we use RDD operator of Spark. RDD is elastic distributed data set and is a core abstract concept of Spark. Simply, first of all, organize all nodes in the cluster. Secondly, enable them to parallel operation. Here, the execution process is divided into three stages (the following Figure):

First stage: data preprocessing:

In this stage, BatMapis operated firstly. This operation is for each row in a data file, the output format should be a key value pair, such as< URL, out link >. The first column content is URL of the current web page, the corresponding value is web page linked byURL. The PageRank values is calculated and reduce By Key operation is performed. In this operation, we should enter a key value pair, which is URL of current web page. The value is a list, and the corresponding first element is PageRank value. The other elements are web pages linked byURL.

The second stage: iteration computation

This stage is mainly for the iterative calculation of processing data, and the iterative processing data is regarded as output data. At first, each output key is each link of input value; it is also called map operation inversion key value pair. However, the output value is links of input key + PR value of input key + the number of external links. Then reduce By Key operation is performed. The new value can be calculated by inversion parameter of value. If the PageRank value is convergence, the third stage can be performed. If the PageRank value is not convergence, using map operation, data is transformed into output formats of the first stage. This is a process of iteration calculation.

The third stage: sorting return

The process of sorting return stage is that the PageRank value is regarded as the key, web links are regarded as values, the format of key value pair is transformed with map operation, then the sort By Key operation can be performed to order from high to low, at last, return to results. The calculation of three stages is in the end.

5.3. Improve the accuracy of approximate solution of PageRank algorithm

This is a natural question: when the difference of approximate solution and the exact solution is the least, is this approximate solution the the best solution? If x is the exact solution of $Ax = b \neq 0$, x is the approximate solution, $r = b - Ax$ is

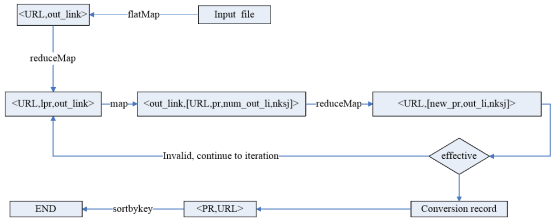

Fig. 3. Algorithm performance process

residual vector, r is very small, is $\bar{x}$ the better approximate solution of $Ax = b$

Theorem 1 (post error estimation): If A is nonsingular matrix, x is the exact solution of $Ax = b \neq 0$, $\bar{x}$ is the approximate solution, $r = b - A\bar{x}$, so the following formula is got:

$$\frac{\left\|x - \bar{x}\right\|}{\|x\|} \leq cond(A)\frac{\|r\|}{\|b\|} \quad (8)$$

Prove: Because of $x - \bar{x} = A^{-1}r$, we can get

$$\left\|x - \bar{x}\right\| \leq \left\|A^{-1}\right\| \cdot \|r\| \quad (5)$$

and $\|b\| = \|Ax\| \leq \|A\| \cdot \|x\|$, so

$$\frac{1}{\|x\|} \leq \frac{\|A\|}{\|b\|} \quad (6)$$

By (9) and (10) , formula (10) is got.

The formula 8 shows that the accuracy (error bound) of approximate solution $\bar{x}$ not only relies on residual$r$, but also relies on the number of conditions of matrix$A$. If matrix A is false, r can no longer guarantee that $\bar{x}$ is highly approximate solution.

In order to improve the accuracy of the approximate solution, according to the condition number of coefficient matrix of the equation group, $K = \|I - dp^T\|_1 \left\|\left(I - dp^T\right)^{-1}\right\|_1 = \frac{1+d}{1-d}$, the result is a serious function of the damping factor d.

The Table 1 is the corresponding conditions number K values which d is in the interval $[0, 1]$ to take 101 values.

Table 1. The corresponding conditions number$K$values that d is in the interval $[0, 1]$

d values	K values	d values	K values	d values	K values	d values	K values	d values	K values
0	1								
0.01	1.2121	0.21	1.5236	0.31	2.8989	0.51	4.0125	0.71	9.2655
0.02	1.0406	0.22	1.5126	0.32	2.4524	0.52	4.2588	0.72	10.2541
0.03	1.2546	0.23	1.5621	0.33	2.5089	0.53	4.2658	0.73	10.2548
0.04	1.8562	0.24	1.6254	0.34	2.5714	0.54	4.2698	0.74	11.2658
0.05	1.5246	0.25	1.6668	0.35	2.6252	0.55	4.2658	0.75	12.2236
0.06	1.5548	0.26	1.7702	0.36	2.3656	0.56	4.2156	0.76	13.2588
0.07	1.5325	0.27	1.7715	0.37	2.3514	0.57	4.0365	0.77	14.3856
0.08	1.2461	0.28	1.7584	0.38	2.8546	0.58	5.2666	0.78	14.2561
0.09	1.0245	0.29	1.8025	0.39	2.8887	0.59	5.2164	0.79	14.2582
0.10	1.2648	0.30	1.8245	0.40	2.5896	0.60	5.2868	0.80	14.2653
0.11	1.0365	0.31	1.8546	0.41	2.0256	0.61	5.8954	0.81	15.2589
0.12	1.2465	0.32	1.9235	0.42	2.8151	0.62	5.6891	0.82	15.6587
0.13	1.2468	0.33	1.9254	0.43	3.2156	0.63	6.2658	0.83	15.2658
0.14	1.2548	0.34	2.1254	0.44	3.2658	0.64	6.2359	0.84	16.2587
0.15	1.3694	0.35	2.2859	0.45	3.2689	0.65	6.3257	0.85	18.2595
0.16	1.2485	0.36	2.6668	0.46	3.2697	0.66	6.2549	0.86	19.3256
0.17	1.2468	0.37	2.8853	0.47	3.2569	0.67	7.2588	0.87	20.3256
0.18	1.2654	0.38	2.3526	0.48	3.8989	0.68	7.5898	0.88	26.2589
0.19	1.4587	0.39	2.2563	0.49	3.9992	0.69	7.4568	0.89	99.2587
0.20	1.5000	0.40	2.6666	0.50	4.0000	0.70	8.0000	0.90	1nf

Table 1 shows that if$d \neq 1$, matrix is "pathological" serious. For$d \leq 0.75$, the condition number$K \leq 12.2236$, so matrix "pathological" is more and more light. If damping factor d is close to 1, web ranking is more fair, at the same time, $d$can not be too large, if d is larger, the approximate solution of the equation group is larger. When the condition number is controlled within 5, the maximum value of d is 0.57<.0.76, the condition number K =4.0365<13.2588, equation set is better. $I - dp^T$ and vector $v^{'}$ has micro perturbation, the influence on accuracy x is less.

5.4. Comparable analysis of PageRank algorithm

In order to scientifically verify the advantages of improved algorithm, the paper uses stable and systematic robustness analysis. Robustness analysis, is used in different fields, and is a evaluation method of verifying model test results. Because the method has the following characters: contextualization, openness, procedural and hierarchical, it is given full play to the function of its own building, choice, identify

and confirmation to deal with complex and changing environment. Here, with a web design company as an example, based on the existing image database, using the 2000 image data, taking the improved model and the initial model results as the standard, the robustness analysis is as follows:

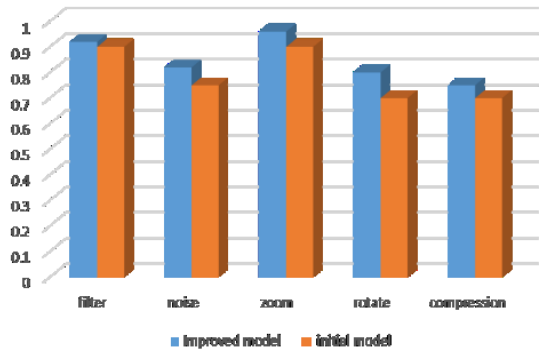

Fig. 4. The robustness analysis of improved model and initial model results

In the above Figure, blue represents the robustness analysis of improved model; orange represents the robustness analysis of initial model. The inspection way respectively discusses five effective conditions: filter, noise, scaling, rotation, and compression, and analyzes the quality of two model graph. Generally speaking, in the changing environment, the quality of improved model graph is higher than the quality of the initial model graph, especially in the rotating environment. And then are noise, scaling and compression. The advantage of filter is not obvious. Therefore, the conclusion is that the application of improved model is beneficial to theory research based on big data background, and when the number of web pages is large, it improves the accuracy of page rank, showing a clear advantage.

6. Conclusion

The improved model is based on optimized insufficiency of algorithm, through the robustness analysis, taking an enterprise image quality as an example; the paper discusses the relationship between the quality of improved model and the quality of initial model, in the different changing environment, such as filtering, noise, scaling and so on. Under the complex background, in the different conditions the quality of improved model and the quality of initial model have advantages. To some extent, the improved model has more development potential in the various backgrounds. Especially in the big data environment, the improved model gets rid of traditional single serial algorithm facing the large data scale and vast amounts of data records. The improved model optimizes the results by the model itself parallelism to meet people needs and society needs.

References

[1] Wu X, Zhu X, Wu G Q: *Data mining with big data.* Knowledge and Data Engineering *26* (2014), No. 1, 97–107.

[2] PAGE L, BRIN S, MOTWANI R: *The PageRank citation ranking: bringing order to the Web.*Proceedings of the 7th 1n-ternational World Wide Web Conference (1998), 161–172.

[3] Wagner A: *Robustness and Evolvability in Living Systems.* Princeton: Princeton University Press (2005).

[4] Wimsatt W C: *Re -engineering Philosophy for Limited Be-ings.* Cambridge: Harvard University Press (2007).

[5] ZHAO L, BAIGSHEN H: *Priority-based IEEE 802.* The Journal of China Universities of Posts and Telecommunications *20* (2013), No. 1, 47–53.

[6] Aouad L M: *Performance Study of Distributed Apriori-like Frequent Item sets Mining.* Knowledge and Information Systems *23* (2010) 55–72.

[7] ZHANG Y, YIN C,WU C: *Research on PageRank algorithm optimization based on Map Reduce.* Application Research of Computers *31* (2014), No. 2, 431–434.

[8] WHITE T: *the definitive guide.* Sebastopol: O Reilly Media (2009), 103–107.

[9] Kraska T, Talwalkar A, Duchi J C: *A Distributed Machine-learning System.*CIDR (2013).

Received November 16, 2017