Research on PageRank Algorithm parallel computing Based on Hadoop

Pengfei Yang
Collection of Information Science and Engineering, Guilin University of Technology, Guilin 541004, China

Liqing Zhou
The modern network technology center, Guilin University of Technology, Guilin 541004, China

ABSTRACT: PageRank algorithm is improved by MapReduce programming model thought based on research and study of PageRank algorithm. MapReduce-based PageRank algorithm run distributed parallel in Hadoop cloud computing platform environments, thereby improving the efficiency of PageRank. The experimental results show that the PageRank algorithm based on MapReduce cluster node number and data block size to affect the efficiency of the algorithm.
KEYWORD: MapReduce model; PageRank algorithm; parallel computing.

1 INTRODUCTION

With the rapid development of the Internet, information of the Internet get more and more rich. Information explosion age has arrived, as of 2015, the scale of China's Internet users have reached 668 million, Internet penetration rate is 48.8%, a large number of Internet users look for information on the Internet, users face so huge amounts of network data information that most of Internet users don't start. How to quickly, efficiently and accurately retrieve useful information, Google founder Sergey Brin and Lawrence Page proposed PageRank algorithm based on link analysis that solve the above problems. MapReduce is a Google cloud-based software framework for parallel processing of large data that can be used for large data sets (more than 1TB) of operation. MapReduce are both functional and vector programming language, MapReduce programming model applies to unstructured and structured mass data retrieval, mining, analysis and machine learning, distributed for large data sets, high computational efficiency(Li Zhiying.2011).

MapReduce computing framework of open source hadoop cloud computing platform, parallel programming can effectively reduce the difficulty and improve programming efficiency, PageRank algorithm is computationally intensive, effective combination of both can improve the efficiency. The paper will examine issues such as PageRank algorithm efficiency under MapReduce.

2 PAGERANK ALGORITHM PARALLEL COMPUTING BASED ON HADOOP

2.1 *PageRank algorithm*

PageRank algorithm was proposed by Google founders Sergey Brin and Lawrence Page and Sergey Brin early in 1997 when the search system prototype proposed link analysis algorithm. The algorithm is based on the algorithm web hyperlink structure analysis of important algorithms and calculated the importance of web pages by link structure for page rank. The core idea of PageRank : When there is a link to the web page A page B, B is considered to obtain the contribution of its A value, depending on how much the value of A's own importance, that the greater the importance of the page A, contribution page B was the higher. Due to mutual links pointing to the page based on the final web page scores that are calculated as an iterative process to search and sort.

PageRank algorithm uses surf model (BooVooi Keong.2011)to calculate page PR value. Configured with the Web Figure G = (V, E), where V is the vertex set of all pages, while E is the set of links between pages, there are links to page A to page B that vertices A, B exists between an edge. F(u) is assumed to be a collection of links to web pages u, B_u of linked pages u collection of pages, so that N (u) = | Fu |, u of the page PR (u):

$$PR(u)= \sum_{V \in B_u} \frac{PR(V)}{N(V)} \quad (1)$$

Formula (1) can be random roaming model to describe the page. When the probability of a user based on the current page to access links to other pages of equal probability, the page PR (u) value is to be accessed the same time when you reach other web browsers continue to follow the link. Therefore a large number of link page or pages to be heavily weighted links to pages with greater weight. formula (1) there are two small problems, one page does not link to other pages, only there is mutual link in a small area, the other is the existence of a out-degree 0. PR value is not distributed out of accumulated and continued cycle, we call this phenomenon weights precipitation. To solve the above problems, Lawrence Page and Sergey Brin of formula (1) has been modified, the user will assume a certain probability to jump to other pages, and continue browsing. Jumping to assume equal probability occur on every page to give the following new formula (2)(White T.2012).

$$PR(u)= \frac{1-d}{n} + d\sum_{v\in B_u}\frac{PR(v)}{N(v)} \quad (2)$$

Where n is the total number of pages, jumping from one page to any other page that damping coefficient is d (general value of 0.85).

2.2 *Hadoop*

Apache Software Foundation's Hadoop is an open source distributed computing platform. Hadoop Distributed File System (HDFS) and MapReduce as the core Hadoop system provides users with low-level details transparent distributed computing framework. HDFS has a high fault tolerance features, and it is designed to be deployed in low hardware. And it provides a high transmission rate to access the application data for those with large data sets applications. HDFS relaxes POSIX requirements of this form of access to the data flow file system. So Hadoop platform hava high reliability, high scalability, fault tolerance, and high efficiency and other advantages (Liu Gang.2014).

Hadoop MapReduce hadoop distributed computing is one of the core of the components of open source frameworks. Hadoop MapReduce parallel programming model is based on Google MapReduce implementation which it is a processing and generating large data sets algorithm model(Jiang Wuxue, Jin.2010).MapReduce design ideas derived from the functional programming language, the expression of a large-scale distributed computing to parallel operation sequence a series of key / value pairs. First users create a map function that process the input data which every logical base on key/value pairs. Then reduce deal with, according to key values appropriately merge operation in the list of elements. The final output result. Its implementation involves many complex details (Xu Yaling.2015). Figure 1 depicts a parallel computing process of MapReduce.

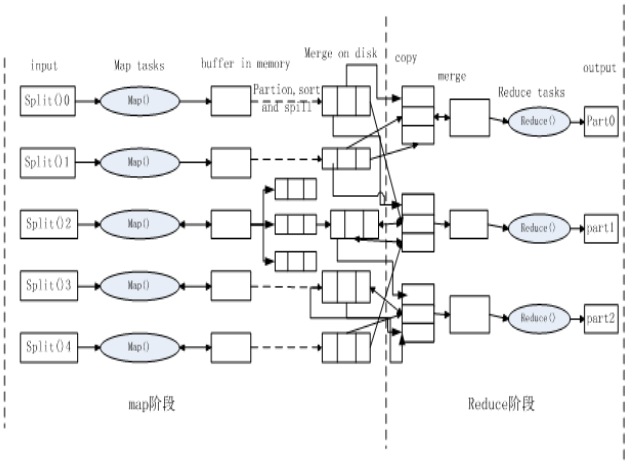

Figure 1 Parallel computing process of MapReduce

2.3 *Implement PageRank algorithm based on MapReduce*

The key of PageRank algorithm is iterative process, the key implementation of PageRank algorithm based on MapReduce is parallel iteration, the information between each of the pages is independent of each other, but each is calculated independently. In the algorithm, each iteration calculates the page of score to use MapReduce processing in order to speed up the efficiency of the algorithm (Yang GaiZhen.2011).

First, the files of link relation store HDFS (distributed file storage system) in Hadoop platform. In storage process, the file is divided into blocks and stored in datanode nodes, while avoiding backup because the file is missing or corrupt files appear downtime Case. Map function process that people divide data blocks, it generates data format<key, value>, and merge data. At last the master calls the Reduce function that processe and produce results, and stored in HDFS. Firstly, data preprocessing, text links relationship convert to links relational tables. The link relational tables store in HDFS, and it is divided into blocks stored on different nodes DataNode then MapReduce operation (Cai Bin, 2013).

Map function

Map function input data<key, value> which page link tables read in HDFS. key representatives of offset, value representatives of the page ID and outlink page ID. Map function read each record sequentially from the text, and save each record to the (key, value) . In order to facilitate consistent and map output stage Reduce the input stage, the key input for this Map of the form <offset, page ID and oulink page ID set>

Map function pseudo-code

Map(key,value)//read a line
//key offset
/value page ID and oulink page ID set
value.split(“,”); // value split page ID and oulink page ID set
for each list in value // text’s record lines
for each outlinklist in value //outlink set
Output(key1,value1);// key1:inlink page ID, value1 page ID,initial pagerank value and outlink page ID set.
Output(key2,value2);// key2 page ID,value2 outlink page ID.
Reduce function
Reduce function manly process that Map function generate data. MapReduce PageRank algorithm,it calculate the final score of each page. its output results is as same as the input of Map function, in ordert that next iteration. Reduce function is output in the form of <page ID and page PR value, outlink page ID set >
Reduce function pseudo-code：
Reduce(key2,values) //key2 page ID, values inlink page
//set,PR value and outlink total quantity
for value in values:
st()= value.split(“;”);//split inlink page information
if st //judge st length
cumulate contribution points
else //record outlink page set
PRpage= cumulate contribution points +(1-a)
output(key3,value) //key3: page ID and PR value
value3: outlink page set.

3 EXPERIMENT

In order to verify the efficiency of MapReduce-based PageRank algorithm, the experiment will use comparative methods to analyze its performance from the amount of data on the efficiency of different sizes.

3.1 *Experiment platform*

In a 2.4GHZ intel (R) Xeon (R) E5645 12-core processor, the memory is 8G, 500GB mechanical hard drive and two 2.4GHZ intel (R) Xeon (R) E5645 12-core processor, the memory is 4G, mechanical hard drive 500GB. Due to limitations of condition, the experiment uses VMware vSphere technology. Using VMware vSphere virtual platform, it generate eight virtual computer include hard disk 100GB and memory 2GB.Built distributed environment, every virtural computer install centos6.6 64byte OS, JDK1.7u71, Hadoop2.2.0 and zookeeper-3.4.5. Among one virtual computer is master node, a virtual computer is master spare machine. a virtual computer yarn host, another are slave nodes.centos1 is Namenode,centos4-centos8 are DataNode. Initial PageRank value is 0.85, PageRank damping coefficient is 0.85. Datasets Stanford University social network research platform provides Web map data. table 1.

Table 1 data set

name	Size (KB)	edge number	node number
Note Dame web graph	21050	1497134	325729
Stanford web graph	32118	2312497	281903
Google web graph	73614	5105039	875713
Berkeley-Stanford web graph	107555	7600595	685230
Patents citation network	273964	16518948	3774768

3.2 *Experiment design*

PageRank algorithm based on MapReduce, first the main impact of the number cluster of Hadoop platform, along with an increase in the number of machines in the cluster, the higher execution rate. the second, the file data is stored in HDFS block size, it is directly related to how much the block have effect on efficiency of Map function. For these two aspects, the first experimental test platform Hadoop clusters increasing in the number on the implementation of the five data set, the second setting different data block size to perform five data sets, finally analyzing the results.

3.3 *Experiment analysis*

PageRank algorithm based on MapReduce running time of data collection on Hadoop platform.

For the different data sets to run six clusters, seven clusters, eight clusters running parallel PageRank algorithm, record iteration phase MapReduce running time, every experiment three times and averaged.as shown in Table 2.

Table 2 different clusters running time(ms)

Cluster node / Web graph edge	Six clusters	seven clusters	eight clusters
1497134	127314	108726	101502
2312497	157986	139697	137845
5105039	369946	330521	327883
7600595	396634	364042	359290
16518948	1251974	1151464	1103682

From Table 2, the number of edges in the web the same situation can be seen with the increase in the number of cluster nodes, PageRank algorithm iteration time is also reduced. Due to an increase in the number of cluster nodes, the cluster processor and memory are fully utilized, thereby improving opera-

tional efficiency. MapReduce computational framework has good expansion.

For data sets are stored in HDSF of 16M, 32M, 64M, 128M data block size to run the program .Data block size calculations have a significant impact on the operation of the data in Hadoop. How much influence the number of data block size affects the data block, the data block Map function of the number of tasks to perform, thus affecting the overall efficiency. Note also that the data block is not the bigger the better, the larger the data block leads to reduction in the number Map function to perform tasks, thus affecting the entire cluster waste of resources. The smaller the data block is not very good, it will result in more Map function performs data, leading to the end of the merge function Map wasting a lot of time, a reasonable block size, thereby improving the overall efficiency.

As shown in Table 3 Figure 2 shows the 16M, 32M, 64M, 128M data block running time: (averaged three times).

Table 3 different block size running time (ms)

Block / Web name	16M	32M	64M	128M
webNotreDame	111223	114390	105969	130618
webStanford	148517	153348	152917	157986
webGoogle	353962	366660	367622	369946
webBerkStan	412723	429467	416105	396634
citPatents	1701398	1178061	1107440	1251974

Table 2 and Figure2can be drawn: the data block based on reasonable size will help improve the efficiency of PageRank algorithm. based on MapReduce.

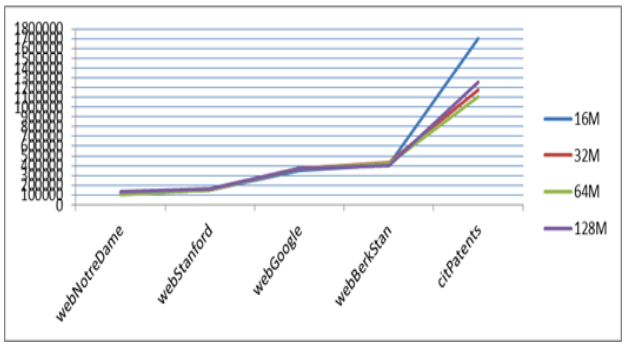

Figure 2 different block size running time(ms)

4 CONCLUSION

The paper analyses and studys parallel Calculation of MapReduce-based PageRank algorithm. while it can effectively integrate PageRank algorithm and MapReduce calculation framework on Hadoop cloud computing platforms, improving the efficiency of PageRank. The use of Hadoop cloud computing platform, the number of test different cluster nodes to perform different datasets on the impact of the algorithm, test change on HDFS data block size on the impact of algorithm performance. Experimental results show that the higher the efficiency, the more the number of cluster nodes to perform reasonable block size will help improve the efficiency of the algorithm. Wasting time communication problem between the experimental ignored cluster nodes, HDFS time access issues, will be among the next step, to proceed deficiencies. Experimental results show that: the more efficiency, the more the number of cluster nodes perform, reasonable block size will help improve the efficiency of the algorithm. Wasting time communication problem between the experimental ignored cluster nodes, HDFS time access issues, etc. I will continue to complete the deficiencies in the future.

REFERENCES

BooVooi Keong.2011. PagePank :A Modified Random Surfer Model. IEEE Conference Publications. 1-6.

Cai Bin, Chen Xiangping.2013.Hadoop Technology Insider:Hadoop Internals:in-depth study of Common and HDFS.China Machine Press.

China internet network information center.2015.The 36st Statistical Survey Report on the Internet Development in China.China Internet.

Jiang Wuxue,Jin.2010.Study on Parallel Programming Framework Model Based on MapReduce.Microelect Ronics & Computer. 27(6):168-170.

Li Zhiying.2011.Research on PageRank. Computer Science.3 8(10).

Liu Gang.2014.Hadoo Application development technology explanation.China Machine Press.

Qian Gongwei.2007. Extended PageRank algorithm based on Web link and content analysis.(J). Computer En.43(21).

White T.2012. Hadoop: The Definitive Guide.3rd ed.(S.1.): O'Reilly Media Inc.

Xu Yaling.2015.Data Placement Strategy for MapReduce Cluster Environment.Journal of Software. 26(8):2056-2073

Yang GaiZhen.2011.The Application of MapReduce in the Cloud Computer(C).IEEE Conference Publication.