ISSN : 0974 - 7435 Volume 10 Issue 5

BioTechnology
An Indian Journal

FULL PAPER

BTAIJ, 10(5), 2014 [1345-1351]

Block matrix-based marpreduce pagerank algorithm web structure mining applied effect research

Weizhong Yan
Department of Computer and Information Engineering, Changzhou Institute of Technology, Changzhou 213111, Jiangsu, (CHINA)
E-mail: 18970447@qq.com

ABSTRACT

Web page not only has text messages, but also contains hyperlinks that points from one page to another one and hyperlinks contain potential annotations. Lots of Web hyperlinks information provides relative Web page contents correlation, quality and structure aspect information, the information reflects documents containment, quotation or affiliation relations. And Web structure mining is mining derived knowledge from World Wide Web organization structure and link relations on Web pages link structures. In information searching, it can regard high authority score and pivot score's webpage as high quality webpage, during searching process, it priority provides it to users, in this way it can discover network community by analyzing hyperlinks' topology and construct a digraph for searching result or assigned webpage set. The paper on the basis of introducing Web structure chart, it analyzes Pagerank algorithm applied merits, and then researches on block matrix-based Mapreduce PageRank algorithm, the method uses block matrix thought to reduce every time iteration mixed phase and rank phase time consumption so that let every time iteration only execute one Mapreduce phase, for the algorithm, the paper compares it with other two algorithms, gets that the algorithm superiority degree on operation time that provides theoretical basis for Web structure mining techniques.

KEYWORDS

Web structure mining;
PageRank algorithm;
Block matrix;
Mapreduce model.

INTRODUCTION

Data mining technique has been rapidly developed since 1990s, theoretical researches have been considerable deepen, and it has been widely applied in all fields, its research range includes association rules mining, classification rules mining, clustering rules mining and trend analysis so on aspects. Meanwhile, in current information world, Internet plays quality roles in information transmitting effects among people, with Internet rapidly development, network has already developed into constantly expanded distributed information space with three hundred million pages, from which covers technical data, commercial information, newspaper report and entertainment information as well as other lots of heterogeneous medium unstructured information, unstruc-

FULL PAPER

tured data nearly covers around 80% of enterprise information sources, and database data only occupies around 20%. Therefore, expand data mining research objects range, it should more focus on unstructured data, such as: text, network page, e-mails and so on, now network mining, text mining and multimedia mining emerge at the right time. In World Wide Web organization structure and link relations, for the text, except for having certain link nodes, World Wide Web can provide useful information beyond documents contents, use these information can rank the pages and discover more important pages, quote documents tend to be more objective, general and accurate on quoted documents explanation, it is helpful for automatic deducing page authority, it can construct a digraph for searching results set, from which every node represents a webpage, nodes directed edges represent hyperlinks, in this way it can generate Web graph, use data mining theory to implement algorithm on modeled Web topology that provides theoretical method for Web structure mining.

The paper on the basis of previous research results, it analyzes Web structure chart and PageRank algorithm, discussed Mapreduce-based PageRank algorithm that provides theoretical basis for Web structure mining, and uses experiment to verify the algorithm validness and accuracy.

WEB STRUCTURE MINING ANALYSIS

Web structure mining is mining Web pages link structures. By analyzing hyperlink topology, it discovers network community, to search results, it can construct a digraph, such structured digraph is called Web graph. In the following, it analyzes Web structure graph concept and link relations.

Web graph structure construction

Web itself has certain stability and quantizable features; therefore use some random selected Web sub set to analyze Web properties is feasible. From the perspective of graph theory, it can regard Web as one digraph that locates in physical network $G(V,E)$, and can abstract understand Web as network graph, from which node set V corresponds to Web webpage, PDF and other documents, and frontier set E corresponds to nodes hyperlinks. Therefore, it constructs Web whole system graph, hyperlink is a bond of information mutual connections, it reflects network information units relations, as Figure 1 showed Web graph.

In Figure1, node that is composed of webpage is

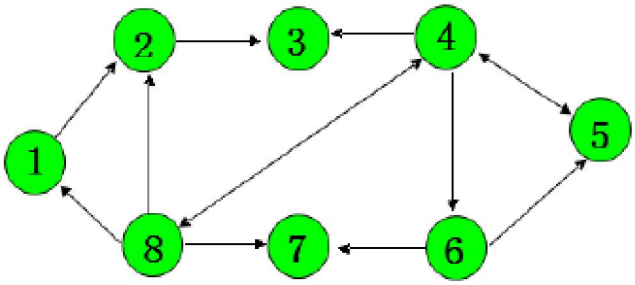

Figure 1: G figure link relations

using set V to express as formula (1) show:

$$\mathbf{V}=\{1,2,3,4,5,6,7,8\} \quad (1)$$

Directed edge set that is composed of webpage hyperlinks is using E to express, as formula (2) show:

$$\mathbf{E}=\begin{Bmatrix}1\to 2,2\to 3,4\to 3,4\to 5,4\to 6,4\to 8,5\to 4,\\ 6\to 5,6\to 7,8\to 7,8\to 2,8\to 1,8\to 4\end{Bmatrix} \quad (2)$$

Stipulate that any two nodes in Figure 1 is $p,q(p\neq q)$, then node p has a piece of hyperlinks that points to q, which represents as $p\to q$, then q is p Outlink webpage, p is q linked page, set $F(p)$ is node p pointed other nodes set, in Figure 1, $F(4)=\{3,5,6,8\}$, set $B(p)$ is pointed node p other nodes set, in Figure 1 $B(4)=\{5,8\}$, define nodes Outlink amount is node out-degree, node Inlink amount is node in-degree, then in Figure1 node 4 out-degree is 4, in-degree is 2.

Link relations matrix construction

Preprocess with Web obtained results and get a URL list, and for every URL, define an ID, when parse Web webpage, parse hyperlink relations, form into link relations data set, as TABLE 1 show the link relations.

According to Figure 1, it lists out TABLE 1 link relations data table, relation represents corresponding ID outlink, regulate when any two nodes have link relations, it represents as 1, when they don't have link relations, it represents as 0, in this way it can get Web graph adjacent matrix expression, as formula (3) shows:

$$\mathbf{G}=\begin{pmatrix}0&1&0&0&0&0&0&0\\0&0&1&0&0&0&0&0\\0&0&0&0&0&0&0&0\\0&0&1&0&1&1&0&1\\0&0&0&1&0&0&0&0\\0&0&0&0&1&0&1&0\\0&0&0&0&0&0&0&0\\1&1&0&1&0&0&1&0\end{pmatrix} \quad (3)$$

TABLE 1 : Link relations data table

URL	ID	relation
http://aaaaaaaaaaaaaa/index.html	1	$1 \to 2$
http://bbbbbbbbbbbbb/index.html	2	$2 \to 3$
http://ccccccccccccccc/index.html	3	0
http://ddddddddddddd/index.html	4	$4 \to 3,4 \to 5,4 \to 6,4 \to 8$
http://eeeeeeeeeeeeee/index.html	5	$5 \to 4$
http://fffffffffffffffffff/index.html	6	$6 \to 5,6 \to 7$
http://ggggggggggggg/index.html	7	
http://hhhhhhhhhhhhh/index.html	8	$8 \to 7,8 \to 2,8 \to 1,8 \to 4$

In formula (3) elements is expressed by every line corresponding outlink existence or not, when exist outlink, it is 1, when don't exist outlink, it is 0, but when node scale is great, adjacent matrix has lots of 0, it is relative difficult to implement algorithm, so in order to save useless storing space, it can equivalently express adjacent matrix as formula (4):

$$\begin{cases} G = \{\alpha, \beta^T\} \\ \alpha = \begin{pmatrix} 1:2 \\ 2:3 \\ 0 \\ 4:3,5,6,8 \\ 5:4 \\ 6:5,7 \\ 0 \\ 8:1,2,4,7 \end{pmatrix}, \beta = \begin{pmatrix} 1:8 \\ 2:1,8 \\ 3:2,4 \\ 4:5,8 \\ 5:4,6 \\ 6:4 \\ 7:6,8 \\ 8:4 \end{pmatrix} \end{cases} \quad (4)$$

In formula (4) α, β^T respectively express corresponding ID outlink and inlink, by formula (4), it is clear that node out-degree and in-degree are as TABLE 2 show:

TABLE 2 : Corresponding node out-degree and in-degree table

p	$Card[F(p)]$	$Card[B(p)]$
1	1	1
2	1	2
3	0	2
4	4	2
5	1	2
6	2	1
7	0	2
8	4	1

PAGERANK ALGORITHM DESIGN ANALYSIS AND IMPROVEMENT

In information searching process, it needs to look for useful information from massive information, therefore for Web structure mining algorithm is particularly important, by far the most classical Web structure algorithm is PageRank algorithm. In the following, analyze the algorithm, and put forward improving opinions.

Algorithm analysis

PageRank algorithm relies on webpage huge link structure to reflect every page quality, the paper takes hyperlink that points webpage p to webpage q as evaluate webpage q authority score, when q in-degree value becomes bigger, then its authority will be higher, of course a webpage, PageRank value not only considers its in-degree value (inlink amount), but also considers in-degree value that points to webpage q webpage p, and so on, it can get PageRank value. Based on above principle, it can get PageRank value (webpage q authority score)decisive factors include:The webpage inlink number and backward chain source page inlink number. Due to a page will point to lots of other webpage; these web pages authority value is equal distributed by the webpage PageRank value points to it. If in Web, webpage number is using n to express then webpage q PageRank value computational

FULL PAPER

method is as formula (5) show:

$$\mathbf{pr(q)} = \sum_{(p,q)\in E} \frac{\mathbf{pr(p)}}{\mathbf{Card[F(p)]}} \quad \mathbf{(5)}$$

In formula(5), $Card[F(p)]$ represents webpage p out-degree, $pr(p)$ represents webpage p PageRank value, $(p,q)\in E$ represents hyperlink that exists $p \to q$, it can regard formula (5) formula as a n pieces of unknown numbers contained n order linear equation set, then pr is using n dimensions' PageRank vector to express as formula (6)show:

$$\mathbf{pr} = [\mathbf{pr(1), pr(2), \cdots, pr(n)}]^{\mathbf{T}} \quad \mathbf{(6)}$$

Calculation principle is as Figure 2 show:

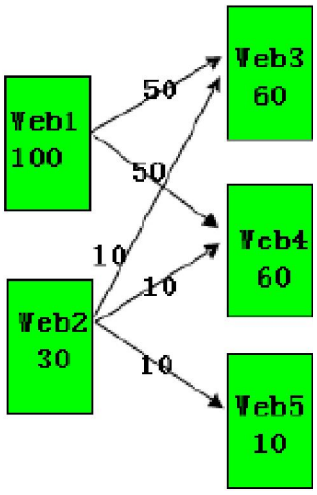

Figure 2 : PageRank algorithm principle

In order to change adjacent matrix each column vector sum after transposition into 1 (total probability), so let each vector compares with respective link number, the rank array is called transposition probability matrix, due to PageRank don't focus on how many places it l inks but how many places are linked, then PageRank matrix can use formula (7) to express:

$$\mathbf{A_{pq}^{T}} = \begin{cases} \dfrac{1}{\mathbf{Card[F(p)]}}, \mathbf{When(p,q)\in E} \\ \mathbf{0, Others} \end{cases} \quad \mathbf{(7)}$$

In PageRank algorithm, if there are web page $P_1 \cdots P_n$ have links that point to webpage A, then set $PR(A), PR(P_1) \cdots PR(P_n)$ respectively to express webpage A, $P_1 \cdots P_n$ PageRank value, parameter d represents a skip coefficient, and $d \in (0,1)$, in the paper it takes $d = 0.85$, $o(A)$ represents A out-degree, then webpage A PageRank value $PR(A)$ expression is as formula (8) show:

$$\mathbf{PR(A)} = (\mathbf{1-d}) + \mathbf{d} * \sum_{i=1}^{n} \frac{\mathbf{PR(P_i)}}{\mathbf{o(P_i)}} \quad \mathbf{(8)}$$

Formula (8) vector computational method is as formula (9) show:

$$\mathbf{P} = \mathbf{dA^{T}P} + \mathbf{(1-d)e} \quad \mathbf{(9)}$$

Traditional PageRank algorithm pseudo code is as Figure 3 show:

PageRank-Iterate (G)

1 $P_0 \leftarrow (1-d)e$

2 $k \leftarrow 1$

3 Repeat

4 $\quad P_k \leftarrow (1-d)e + dA^T P_{k-1}$

5 $\quad k \leftarrow k+1$

6 util $\|P_{k+1} - P_k\| < \varepsilon$

7 return P_k

Figure 3 : Traditional PageRank algorithm pseudo code

Algorithm convergence estimation

In ideal state, PageRank algorithm needs to iterate n times n is number of Web nodes. But in actual application process, it doesn't need to speed long time to iterate so many times, as long as PageRank value is in a reasonable range then it can arrive at user demand, the paper applies formula (10)as judgment algorithm convergent criterion to reflect Web graph convergence speed:

$$\frac{\|\mathbf{P_k - P_{k-1}}\|}{\mathbf{n}} < \boldsymbol{\varepsilon} \quad \mathbf{(10)}$$

And set that in formula (10) $\varepsilon = 0.000001$, when meet formula (10), we think algorithm arrives at convergence. Therefore judge a Web graph convergence speed, firstly is focusing on edge number and node number ratio, the ratio gets bigger and convergence speed will get slower.

Algorithm deficiency analysis and improvement

PageRank algorithm is calculating page authority

algorithm by off-line way, the purpose is to get authority score quantized value, therefore in information searching process, it has faster reaction capacity, to the algorithm, we should more focus on the problems that in one hand is to improve judgment webpage authority accuracy rate, in the other hand, it is to improve algorithm executing speed and reduction memory consumption.

Deficiency

When webpage amount is extremely big, it will lead to node numbers become more, and also causes Web graph adjacent matrix magnified, Web graph adjacent matrix mostly is sparse matrix, now lots of storing resources and calculation resources are vacantly occupied.

Improvement

With internet constantly development, webpage information data amount has already become massive, save and calculate webpage link relations also need to be implemented through large-scale parallel system, therefore PageRank algorithm parallelization development becomes necessary, MapReduce-based PageRank algorithm emerges at the right moment.

MAPREDUCE-BASED PAGERANK ALGORITHM

MapReduce programming principle

MapReduce frame uses map and reduce two functions to implement data cutting and merging, MapReduce model has an advantage of high abstraction, the model includes Map function,Reduce function and <key, value> pair these three core parts.

***Map Function**

User defined Map function receives input data slice generated <key, value> matching, then generate a series of medium <key, value> matching, MapReduce library respectively polymerizes every possessed same medium key medium value and sends to Reduce function.

***Reduce function**

Reduce function receives a medium key and its corresponding value set, meanwhile provide the key value set in the way of integrator for Reduce function, now user defined Reduce function summarizes these value and outputs them. Reduce generally is used to summarize data, programs large scale data and gets smaller data, such as: Implement a"+" operation that is returning input data value list sum.

***<key, value> pair**

<key, value> pair is composed of value representative task related data and key representative value grouping codes two parts, value needs to be put in the corresponding groups to participate in calculation process. Therefore, <key, value> pair can be regarded as programmer supplied communication interface.

Map task and Reduce task is a entirety, inseparable. In one time MapReduce process, Map tasks are in parallel, Reduce tasks in parallel, and Map task and Reduce task are in serial, one time MapReduce process is in serial with next time MapReduce process, these operations synchronization is ensured by MapReduce system. In addition, MapReduce can also express following five kinds of problems.

Distributed searching

Map function according to assigned mode to match to specific line, and transfer it to Reduce function, Reduce function takes these medium results as final result output ;

URL access frequency statistics: Map function handles with webpage request log, output<URL, 1>.Reduce functions makes accumulation on same URL, combine them into <URL, Total> pair;

Inverted index: Map function analyzes every document, and then generates a (Word, Document NO.) pair sequence. Reduce function received a given word all pairs, ranks corresponding documents IDs and generates a(Word, ID List)pair, all output matching forms into a simple inverted index, in this way it can simply increase this position tracking calculation;Host term vector: a term vector uses a(Word, Frequency)list to summarize most important words that appear in a document or a document set. Map function is that every input document generating a (Host name, Term vector) pair. Reduce function receives given host all documents' term vector, and add these terms vectors together, get rid of low frequency term, and finally generate a (Host

FULL PAPER

name, Term vector)pair;

Reverse network connection graph: Map function outputs (Target, Source) pair to every link, every URL is called Target, page that contains the URL is called Source, Reduce function according to give correlative target URLs connects all sources URLs and forms into a list, and generates(Target, Source list)pair;

Block matrix principle

Due to PageRank algorithm is a iterative calculation process, after iterating every time, it is saved in memory that is used to next time iteration, so parallelization, PageRank algorithm every time iteration is a MapReduce calculation process. Block matrix multiplication thought solves slow operation problems when transition matrix size surpasses main memory, as Figure 4 show block matrix multiplication thought.

In Figure 4, according to block matrix thought, set block size as 2, $B_{i,j}$ represents a matrix block, V_i represents a vector block, then matrix multiplication changes into new vector by compounding after every vector block and its corresponding matrix block multiplying, the algorithm can implement parallel effects.

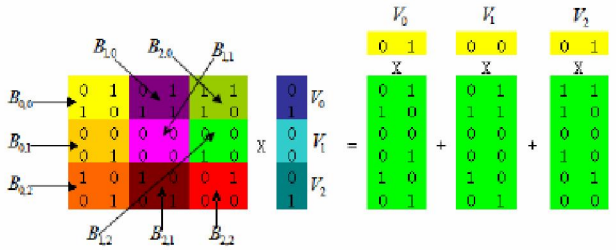

Figure 4 : Block matrix multiplication thought

Web data set data block classification

MapReduce phase 1 function is classifying link relations expressed Web data set into matrix block form, firstly read TABLE 1 link relations and take them as input, use setting block size to reduce total number of blocks, index in every block is equal to every webpage node. Total vector block number is equal to *(nodeNum-1)/blockSize*, every block maximum index number is *(nodeNum-1)modblockSize*, in TABLE 3 showed link relations blocking expression.

In MapReduce phase1, Map method input and output are as TABLE 4 show.

In MapReduce phase1, Reduce method input and output are as TABLE 5 show.

TABLE 4: Map method input and output table

Input	*Page_i Page_j*
Output	Key:*block_id* Value:*in_block_id(page_i) page_j*

TABLE 5 : Reduce method input and output table

Input	**Key:*block_id*** **Value:list[*in_block_id(page_i) page_j*]**
Output	Key:*block_id* Value:PageRank(*i*), Links[y_1, ..., y_m] ...PageRank(*i*+*blockSize*-1), nextlinks[y_1, ..., y_m]...

Block matrix thought-based parallel PageRank algorithm

Iterative phase Map method is reading phase 1Reduce output from HDFS, index in every node block is [(*page/blockSize*)-1], Iterative phase Reduce reads Map output list, maximum block size is *nodeNum/blockSize*-1, index value in maximum block is *nodeNum* %/*blockSize*-1.

Mapper reads initial data and generates partial PageRank value for every block index point, Reducer combines all parts PageRank value for every block index point, and generates new PageRank vector result set, in algorithm template effect is used to record original value, before Reduce process ending, assembles every block every index PageRank value and original record to use for next time iteration. After iteration ending, parse block record into every node PageRank vector structure set, it can get result.

Block matrix algorithm calculation speed contrast

In MapReduce, there are three kinds of PageRank algorithms that are respectively MapReduce PageRank algorithm, low iteration algorithm and the paper's block MapReduce PageRank algorithm, TABLE 6 reflects three algorithms contrast on the same task executing time.

As Figure 5 show three algorithms trend figure.

In Figure 5, green line represents blocking MapReduce PageRank executing time change trend graph as Web edges increase, pink line and yellow line respectively represent MapReduce PageRank executing time and low iteration algorithm executing time change trend followed by Web edges, by Figure 5, it is clear

TABLE 6 : Three algorithms executing time on the same task

Edges time algorithm	Blocking MapReduce PageRank executing time	MapReduce PageRank executing time	Low iteration algorithm executing time
235 ten thousand	75	79	100
505 ten thousand	100	110	135
750 ten thousand	120	131	169
3.5 ten million	1101	1199	1502

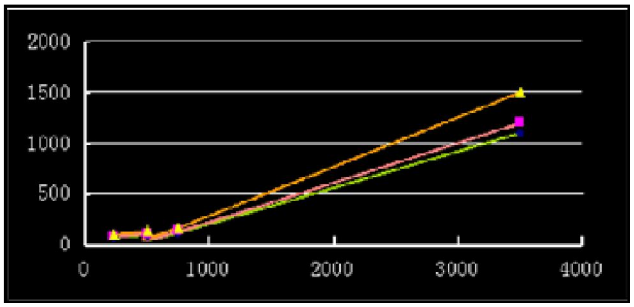

Figure 5 : Calculation time control chart

that blocking MapReduce PageRank executing time has higher superiorities by comparing with other two.

CONCLUSION

Block matrix thought is successfully applied into Web structure mining technology, and it has higher superiorities by comparing with other algorithms; the paper provided algorithm pseudo code can be implemented in general programming software; information exchange plays crucial roles in social development, and Web structure mining technology has higher constraints on information searching effects.

REFERENCES

[1] G.Salton, A.Wong, C.S.Yang; A Vector Model for Automatic Indexing. Communication of the ACM, **18 (11),** 613- 620 **(1975)**.

[2] Keli Chen, Chengqing Zong; A New-Weighting Algorithm for Linear Classifier. Beijing: International Conference on Natural Language Processing and Knowledge Engineering, 650- 655 **(2003)**.

[3] Y.Yang, J.O.Pedersen; A Comparative Study on Feature Selection in Text Categorization. The 14th International Conference on Machine Learning, San Francisco: Morgan Kaufmann Publishers, 412- 420 **(1997)**.

[4] Monica Rogati, Yiming Yang; High-Performing Feature Selection for Text Classification. CIKM'02, New York: ACM Press, 4-9 **(2002)**.