

Comparative Analysis of Web PageRank Algorithm using DFS and BFS Crawling

Poonam Kumari¹, Gaurav Kakhani²

¹ Mewar University, Chittorgarh, Department of Computer Science of Engineering, NH-79, Gangrar-312901

² Co- Guide Mewar University, Chittorgarh, Department of Computer Science of Engineering, NH-79, Gangrar-312901

Abstract: *With the rapid growth of the Web, the amount of information provided over internet is huge, diverse and dynamic. Most of the people use the internet for retrieving information. This valuable information is retrieved by Information Retrieval (IR) Systems. Search engines use web crawling to collect web pages. IR systems use link structure methods which show that the connecting edges (hyperlinks) of web pages give valuable information. To solve a query IR system first crawl web pages to turn unstructured data to structured data. And to crawl it uses various algorithms like Breadth First Crawling, Depth First Crawling algorithms. Here we analyze these two algorithms in real world scenario and compare the results to find out which one gives efficient search results and to find out future scope of research.*

Keywords: Web Crawling, PageRank, Depth-first crawling, Breadth-first crawling, Information Retrieval System.

1. Introduction

World Wide Web contains vast amount of information in unstructured (natural language text) form and provides an access to it at any place at any time. Information Retrieval (IR) [1] system plays vital role to deal with huge amount of data present over World Wide Web in different form such as text, audio, video, images etc. The major challenge of IR system [2] is as given below;

- How to convert unstructured data to structure data.
- How to relate rapidly growing, in number, document.
- How to discriminate documents.
- Relevance of documents.
- How to assess the quality of result.

However, after doing conversion of unstructured data to structured data the next most important job is how to relate these rapidly growing documents and how to assign rank value, page ranking is done to assess the quality and popularity of web page, to them.

1.1 Web Crawling

Web crawling [3] is an approach for converting unstructured data to structured data. It exploits the characteristic of HTML structure, such as metadata; anchor etc., for gathering information. Two approaches Depth-first crawling and Breadth-first crawling is used for web crawling. Due to limitation of hardware, required bandwidth and other network limitations, a web crawler can not download all the pages. It is important to get the order of crawling to get most efficient result and avoid crawling many irrelevant pages.

1.2 Page Rank

PageRank [4] is defined as “It is a measurement of popularity of a web page”. It is a link analysis algorithm used to determine relative importance of a website [5]. It is a static measure designed to rank web pages in the absence of any

queries, means it computes the “global worth” of each page [6]. There are different ways to calculate page rank value and depending on the approach chosen for page rank calculation, page rank value may differ. In link based approach hyperlink to a page is considered as a vote of support and it increases its popularity. The PageRank [7] of a page is defined recursively and depends on the number and page rank metric of all pages that link to it (Inbound links). A page that is linked to by many pages with high page rank value receives a high rank value itself. PageRank [7] is named after Larry Page and used by Google Web Search Engine that assigns a numerical weighting to web pages.

Assume page A has pages T1...Tn which point to it (i.e., are citations). The parameter d is a damping factor which can be set between 0 and 1. Google usually set d to 0.85.

$$PR(A) = (1-d) + d(PR(T1)/C(T1) + \dots + PR(Tn)/C(Tn))$$

Where,

- PR(A) is the PageRank [1] of page A.
- PR(Ti) is the PageRank [1] of pages Ti which link to page A.
- C(Ti) is the number of out links on page Ti and
- d is a damping factor which can be set between 0 and 1.

So, first thing to notice that, PageRank [7] does not rank web sites as a whole, rather it determines PageRank for each page individually, order of crawling a page doesn't affect PageRank. Further, the PageRank of a page A is recursively defined by the PageRank's of those pages which link to page A.

2. Literature Survey

Web crawling plays a vital role to extract required information from huge amount of data. Web crawling should be efficient so that it can provide good results. Various web crawling algorithms has been proposed like Breadth First Search, Depth First Search, Page Rank algorithm, Path-Ascending Crawling algorithm, Focused crawling Algorithm,

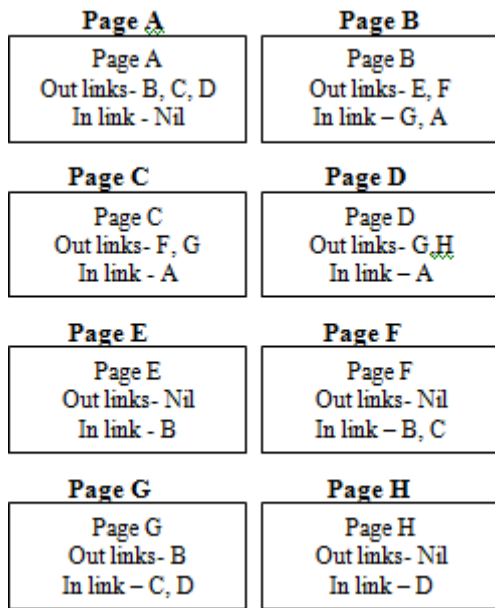
Online Page Importance Calculation Algorithm, Navie Bayes Classification Algorithm and Semantic Web Crawler Algorithm. All these algorithms have their merits and demerits in different scenario. [8]

3. Problem Definition

User has limited and precious time to get information because if user does not get information on time and in accurate manner then that information may become useless for that user. So to give best results to user Search Engines Engineers always try to apply best efficient algorithm. Here we are considering two algorithms Breadth First Search and Depth First Search crawling and by analyzing these algorithms for the real world case study we find out which algorithm gives better results.

4. Methodology

If we consider web as a graph then all the web pages considered as vertices of the graph and all the hyperlinks as the connecting edges of the graph. Here we are taking an example of a website which contains 8 web pages named A, B, C, D, E, F, G, and H. These pages are linked to each other as:



We are assuming page “A” as seed page then we can show the relation among these pages (A, B, C, D, E, F, G, H) graphically as in figure 4.1.

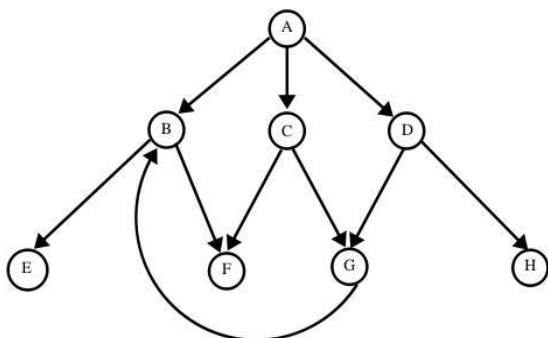


Figure 4.1: hypothetical web graph

When we apply PageRank Algorithm [7] then PageRank retrieved by each page is as follows:

Table 1: PageRank of all the pages in graph

Page	PageRank
PR(A)	0.018750000000000003
PR(C)	0.024062500000000004
PR(B)	0.057385156250000001
PR(E)	0.04313869140625001
PR(D)	0.024062500000000004
PR(G)	0.039203125000000005
PR(F)	0.05336525390625001
PR(H)	0.028976562500000004

Before designing the web crawler, we must have some stop condition on the crawler otherwise our crawler will run for lifetime, because millions of web pages is getting added. Here are the two approaches towards web crawling:

- Breadth First Crawling
- Depth First Crawling

4.1. Depth First Crawling

In Depth First Crawling, our approach is similar to depth first search [9] of a tree or graph. We will start with seed page and crawl deeper and deeper until all the pages on that path get crawled and then backtrack and crawl other branches of the graph. That means as we crawl web pages, we will look at first link on each page in the chain of pages until we get to the end. Only then we will start to look at the second link on the first page and subsequent page. The order of depth first crawling for hypothetical graph depicted above in figure 4.1 is:

A → D → H → G → B → E → C → F

Flow chart of the rank retrieval system using DFS Crawler is as below in figure 4.2:

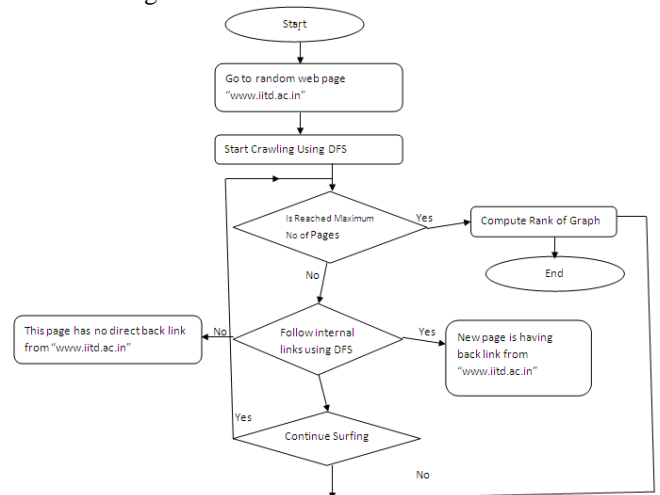


Figure 4.2: Flow Chart of rank retrieval system using DFS crawling

If our goal is to get a good corpus of the web, doing a depth first crawling is the probably not the best way to do that. If

we complete our crawling, no matter what order we follow, we'll find the same set of pages. If we are not able to complete the crawling and with a real web crawler there is no predefined condition to stop the web crawler, so in this case order of crawling matters a lot.

Our approach towards depth first crawling is to provide the crawler a number, total no of web pages to crawl, in advance i.e. after crawling that much web pages our web crawler will stop the crawling new web pages and we are left with limited corpus. However in this approach, we can't predict the order of crawling. It will also affect the ranking of web pages and search result quality.

4.2. Breadth First Crawling

The major disadvantage of Depth First crawling is how to predetermine the end condition and total no of web pages to crawl to produce a good corpus. However, after analyzing the web structure we can say that if we represent web as a tree then height of the tree is important factor to examine so we can fix the depth of the searching starting from the seed pages. This approach does not going to crawl web in one direction. It will crawl seed pages and save all the links whose depth in one more then seed pages and after that it will crawl next level pages. Crawling in Breadth First is similar to searching in Breadth First Search [10].

The order of Breadth First crawling for the hypothetical graph depicted above in figure 4.1 is:

A → B → C → D → E → F → G → H

Flow chart of the rank retrieval system using Breadth First Crawling is as below in figure 4.3:

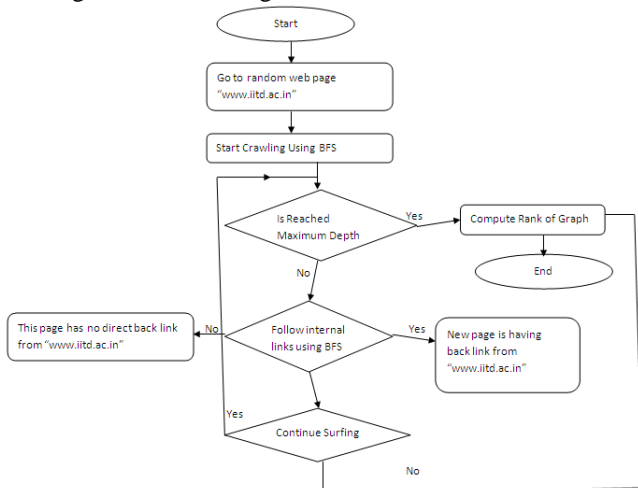


Figure 4.2: Flow Chart of rank retrieval system using BFS crawling

In this approach we start with a constant number, which represent the depth up to which crawler will run. If we consider that upper level pages of a website are of good quality i.e. they are considered as more important than lower level pages and then we are not going to miss them from getting crawl. The major difference between Depth First Crawler and Breadth First Crawler is at the beginning of a program. We can fix the total no of web pages to crawl but can not expect good corpus in Depth First Crawling. Its worst case may happen that one of our seed page did not get crawled rather in breadth first crawling this case will never arise. Each page get crawled at least to given depth even we can determine the depth up to which we have to run our crawler if we do little more analysis on web structure.

5. Result Analysis

We have used Python 2.7 integrated framework to implement Crawling and Page Rank algorithms.

Standard Result

We are considering Google Chrome page rank checker toolbar [11] as a standard result, and we are considering the relative rank. Here is the standard result of few pages of Indian Institute of Technology, Delhi.

Table 2: Standard Result of PageRank for IIT Delhi

Page	PageRank
"http://www.iitd.ac.in"	8/10
"http://nano.iitd.ac.in"	7/10
"http://www.fitt-iitd.org"	7/10

Breadth-first Crawling Result

We run the breadth first crawler at different maximum depth of crawling and note the result and compare it with standard result to check relative order of page rank.

Assumptions: we have few assumptions before starting crawling, our seed page is http://www.iitd.ac.in. We run the crawler at different depth.

Table 3: BFS PageRank Result for IIT Delhi

Page	Page Rank at				
	Depth=1	Depth=2	Depth=3	Depth=4	Depth=5
"http://www.iitd.ac.in"	0.000404028480494	0.00280428851575	0.00215300032557	0.00280428851575	0.00281528851576
"http://nano.iitd.ac.in"	0.000173914497978	0.000500555259217	0.000244224539005	0.000500555259217	0.000500555259217
"http://www.fitt-iitd.org"	0.00018207842946	0.000513155387665	0.000246272081267	0.000513155387665	0.000513165387667

The relative order of page rank for these three pages is same as standard result. Hence, we can conclude that our approach is in right direction.

Depth-first Crawling Result

The limitation of depth first crawler is that to predefine the total number of web page to crawl. We run the depth first

crawler different times by taking different number of web page to crawl and note the result and further we compare it with standard result. The total number of web page to crawl is chosen is the total number of web pages are get crawled by breadth first crawler at different level to ensure the correct comparison of results.

Table 4: DFS PageRank result for IIT Delhi

Page	Page Rank at total no. of pages crawled			
	Page Crawled=500	Page Crawled=1500	Page Crawled=2500	Page Crawled=4500
"http://www.iitd.ac.in"	1.64141041183e-05	5.78059578545e-05	6.71776103875e-05	9.49645982357e-05
"http://nano.iitd.ac.in"	Not Crawled	Not Crawled	Not Crawled	1.33276489367e-05
"http://www.fitt-iitd.org"	1.63313495608e-05	5.75145190755e-05	6.653789358648e-05	8.838757821357e-05

6. Conclusion

Here we have used Google page rank algorithm to find out the page rank of web pages and get the crawling results on the basis of Breadth-first and Depth-first crawling approaches. Results clearly shows that Breadth-first crawling approach gives good corpus and there is always a possibility to get the required page but in case of Depth-first crawling we can not ensure good corpus and in worst case it may happen that one of our seed page did not get crawled. So we can say that Breadth-first crawling gives better ordered result of crawling.

7. Future Work

In proposed crawling techniques we have not discuss about higher level crawling for image and video. In the future we will implement higher level crawling for image and video to minimize fraudulent act. We can modify page rank algorithm by parting the web in various pages like relevant, most relevant, not relevant and extremely not relevant. Finally we need to deploy the proposed technique on internet to serve the internet user.

8. Acknowledgments

Foremost, I would like to express my sincere gratitude to my advisor Mr. Gaurav Kakhani for the continuous support of my study, for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped me in all the time of research and writing of this paper. I could not have imagined having a better advisor and mentor as a student. Besides my advisor, I would like to thank the rest of my Department professors or lecturers or students: Assistant professor B.L. pal, Shiv Kumar and Rohit maheswari, for their encouragement, insightful comments, and hard questions. Last but not the least; I would like to thank our Shiv sir best friend ,Sneha Rani supporting me spiritually throughout my life.

References

- [1] Diana Inkpen, "Information Retrieval on the Internet", Assistant Professor, University of Ottawa, Canada, KIN6N5. (journal style)
- [2] Chris Manning, Pandu Nayak, and Prabhakar Raghavan, *Information Retrieval and Web Search*, Computer Science Department, Stanford University, Stanford, CA 94305, USA. (journal style)
- [3] Eric A. Brewer, *Combining Systems and Databases: A Search Engine Retrospective*, University of California at Berkeley
- [4] Phyu Thwe, "Proposed Approach for Web Page Access Prediction Using Popularity and Similarity Based Page Rank Algorithm" IJSTR volume 2, issue 3, March 2013. (journal style)
- [5] PageRank," <http://web.eecs.umich.edu/~mozafari/fall2014/eecs584/reviews/summaries/summary26.html>". (online content)
- [6] Matthew Richardson, Amit Prakash, Eric Brill, "Beyond PageRank: Machine Learning for Static Ranking" Microsoft Research One Microsoft Way Redmond, WA 98052
- [7] Sergey Brin and Lawrence Page, The anatomy of a Large – scale Hypertextual Web Search Engine, Computer Science Department, Stanford University, Stanford, CA94305, USA.. (journal style)
- [8] Rashmi Janbandhu, Prashant Dahiwal, M.M. Raghuvanshi, "Analysis of Web Crawling Algorithms", IJRITCC, Volmue:2 Issue:3, March 2014
- [9] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein, *Introduction to Algorithms*, Second Edition, MIT Press and McGraw-Hill, 2001. ISBN 0-262-03293-7. Section 22.3: Depth-first search, pp. 540–549(book chapter style)
- [10] Ms. Avinash Kaur, Ms. Purva Sharma, Ms. Apurva Verma, "A Appraisal paper on Breadth First Search, Depth First Search and Red Black Tree", IJSRP, Volume 4, Issue 3, March 2014.
- [11] Google Chrome Page Rank Checker Toolbar, "https://chrome.google.com/webstore/detail/open-seo-statsformerly-pa/hbdkkcheckcdppiaiabobmennhijkkn" [Online]

Author Profile



Poonam Kumari is currently pursuing masters degree program in Computer science and engineering in Mewar University, India



Gaurav Kakhani received the M. Tech. degree in Computer Science and Engineering from Mewar University Chittorgarh. He is working in Mewar University, Chittorgarh as Assistant Professor since 2009.