IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 5, May 2021

DOI 10.17148/IJARCCE.2021.105142

PageRank Algorithm: A Survey

Suraj Singh[#1*], Pranav Shetty[#2]

[#]Department of Computer Science, All India Shree Shivaji Memorial Society's College of Engineering,

Savitribai Phule Pune University, Pune, Maharashtra, India

Abstract: In this rapidly evolving hyper structure, finding relevant information is incredibly challenging. Basically, the aim of a website is to provide useful information to meet the needs of its visitors. As a result, it's critical to locate the content of web pages and recover them based on the user's interest and actions. The important parameter for finding related information is link analysis. PageRank is a major topic among Search Engine Optimization (SEO) professionals. PageRank is based on a mathematical formula that, while intimidating at first glance, is actually very easy to comprehend. In this paper we will be surveying normal implementation of PageRank algorithm with an improved PageRank algorithm (IPRA). The Improved PageRank Algorithm (IPRA), which is an extension of the PageRank algorithm. IPRA distributes the rank score based on the reachability of the web pages, taking into account the number of separate domains in-links and out-links. PageRank and HITS are two fundamental web structure mining algorithms. Based on these two basic algorithms, a plethora of algorithms have been created to improve performance.

Keywords: Mining, PageRank algorithm, Web Mining, Search engines, Websites, In-links, Out-links.

I.INTRODUCTION

One of the tools Google uses to assess the significance or value of a website is PageRank. When it comes to the Google listing, it's just one part of the picture; the other aspects are covered elsewhere (and are constantly changing), and PageRank is significant enough to justify its own article. PageRank is a link analysis algorithm that gives each element of a hyperlinked collection of documents, such as the World Wide Web, a numerical weighting in order to "measure" its relative value within the set. Any list of entities with reciprocal quotations and references can be used with the algorithm. Jon Kleinberg created the Hyperlink Induced Topic Search (HITS) Algorithm, which is a Link Analysis Algorithm that scores webpages. This algorithm is applied to web link structures in order to find and rank related webpages for a given search.

It is very difficult to find relevant information in this rapidly growing web and with widespread use of the web in e-commerce, e-learning, and e-news. As a result, web mining is extremely significant. Web mining is critical for extracting important and potentially valuable information from web data. Web content mining, web structure mining, and web usage mining are the three main types of web mining. The three categories mentioned above are interdependent. Web structure mining is closely related to Web content mining, and the two are linked to Web usage mining. Web structure mining's main goal is to derive information from the web's hyperlink structure. The aim of web structure mining is to examine the link structure of web pages' hyper link structure. Web structure mining's most important parameter is in-links and out-links.

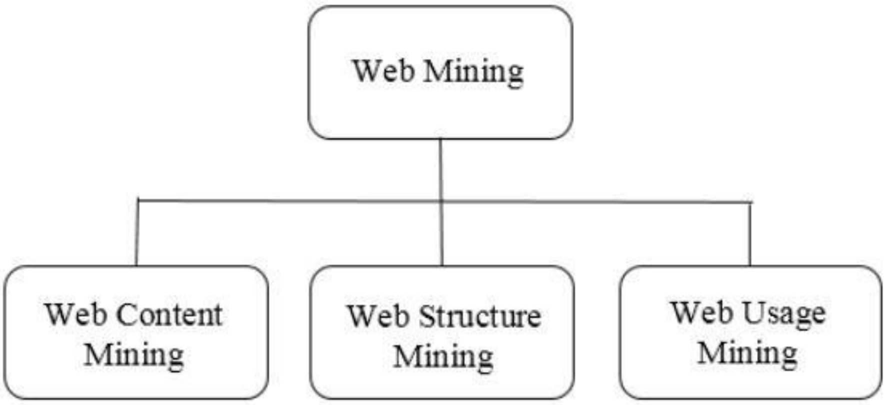

Fig. 1: Web Mining

In this paper we will be stating and comparing two versions of PageRank algorithm the first one is the standard PageRank algorithm and the second one is improved PageRank algorithm. We will survey papers of these two versions.

II.HISTORY OF PAGERANK ALGORITHM

PageRank was created in 1996 at Stanford University as part of a research project on a new type of search engine by Larry Page and Sergey Brin. Sergey Brin proposed that information on the internet could be sorted into a hierarchy based on "link popularity," with a page ranking higher as more links point to it. Scott Hassan and Alan Steremberg, both of whom were cited by Page and Brin as being crucial to the growth of Google, assisted in the development of the method. The first paper about the project, describing PageRank and the initial prototype of the Google search engine, was co-authored by Rajeev Motwani and Terry Winograd with Page and Brin and published in 1998. Page and Brin created Google Inc., the company behind the Google search engine, not long after. PageRank remains the foundation for all of Google's web-search tools, despite being only one of several variables that decide the ranking of Google search results. The word "PageRank" is a reference on Larry Page, the creator of Google, as well as the idea of a web page ranking.

III.PAGERANK ALGORITHM

The PageRank algorithm generates a probability distribution that is used to reflect the possibility of a random individual clicking on links ending up on a specific page. PageRank can be measured for any size range of documents. Several research papers presume that at the start of the statistical process, the distribution is uniformly distributed among all documents in the set. The PageRank computations necessitate multiple passes through the array, referred to as "iterations," in order to change estimated PageRank values to more closely match the theoretical true value. A probability is a number between 0 and 1 that represents the likelihood of anything happening. A "50 percent chance" of anything occurring is generally expressed as a 0.5 probability. As a result, a document with a PageRank of 0.5 ensures that a person clicking on a random connection would be led to that document 50% of the time.

In a nutshell, PageRank is a “vote” about how significant a page is from all other pages on the Internet [3]. A vote of support is represented by a connection to a website. There is no help if there is no connect.

PageRank is described as follows, according to the original Google paper:

We 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. We usually set d to 0.85. There are more details about d in the next section. Also C(A) is defined as the number of links going out of page A. The PageRank of a page A is given as follows:

PR(A) = (1-d) + d (PR(T1)/C(T1) + ... + PR(Tn)/C(Tn))

Note that the PageRanks form a probability distribution over web pages, so the sum of all web pages' PageRanks will be one.

PageRank or PR(A) can be calculated using a simple iterative algorithm, and corresponds to the principal eigenvector of the normalized link matrix of the web.

Let’s break down this method to understand it efficiently:

i. **PR(Tn)** - Every page has its own sense of self-importance. That is, “PR(T1)” for the first page till “PR(Tn)” for the last page of the site.

ii. **C(Tn)** - Each page's vote is equally distributed among all of its outgoing links. “C(T1)” represents the number of outgoing links on page 1, “C(Tn)” represents the number of outgoing links on page 2, and so forth for all pages.

iii. **PR(Tn)/C(Tn) -** if our page (page A) has a backlink from page "n," page A will receive a "PR(Tn)/C(Tn)" share of the vote.

iv. **d(...)** - All of these fractions of votes are added together, but this total vote is "damped down" by multiplying it by 0.85 (the element "d") to prevent the other pages from getting too much control.

v. **(1 – d)** - The (1 – d) bit at the start is a bit of probability math magic that ensures that the "number of all web pages' PageRanks will be one": it replaces the bit missed by the d(.... It also ensures that even though a page has no links pointing to it (no backlinks), it would always receive a PR of 0.15 (i.e. 1 – 0.85).

So scores are computed at indexing time in this algorithm, and the results are sorted by page importance. A page P's Page Rank value is:

$$PR(P) = (1 - d) + d * (\frac{PR(N1)}{C(N1)} + \cdots + \frac{PR(Nn)}{C(Nn)})$$

Fig. 2: Formula of PageRank

Now that we understand the working of standard PageRank algorithm lets now understand the working of the improved version.

IV.IMPROVED PAGERANK ALGORITHM

The Google search engine uses the PageRank algorithm. The Google PageRank algorithm is based on the number of inbound and outbound connections, which is something that almost every website owner is aware of. As a result, they increase the number of in-links and out-links to boost their web page rankings, a process known as mutual reinforcement.

Mutual reinforcement refers to the creation of a bond between web pages in order to boost their page rank. There are a few parameters that can be used to minimise the mutual reinforcing effect, such as the number of in-links and out-links from different domains and reachability. To reduce the mutual reinforcement effect, the improved PageRank Algorithm employs all three parameters.

By using the reachability value of a web page, the proposed Improved PageRank Algorithm, an expanded version of the PageRank algorithm, reduces mutual reinforcement effect [1]. If a web page has several in-links and out-links from the same domain, we have chosen the link(v,u) with the highest reachability and discarded all other links based on reachability value (v,u).

PR(v) is the likelihood of reaching the web page v, which is determined using the formula below.

$$PR(v) = \frac{Total\ number\ of\ reachable\ web\ pages\ of\ web\ page\ v}{Total\ number\ of\ web\ pages}$$

Fig. 3: Formula of Improved PageRank Algorithm

The improved version uses three factors to boost the performance of the PageRank algorithm, by reducing the mutual reinforcement. As now we understand improved version of PageRank lets see HITS algorithm which aids the improved PageRank algorithm.

V.HITS ALGORITHM

To obtain refined pages for a subject query, Kleniberge suggested using the HITS algorithm. He had introduced two interesting words. The first is Hub, which are websites that point to a large number of hyperlinks, and the second is Authority, which are websites that point to a large number of hyper-links. HITS is entirely dependent on relation structure. While in iterations, this algorithm calculates Hub and Authority Scores for relevant sites.

VI.DISCUSSION

In this section, we will discuss, what most of the related works mentioned above provide to us and what were the drawbacks in the schemes proposed by these papers. Web mining is used to collect information from web data. In this method, web structure mining is critical. In web structure mining, HITS and PageRank are widely used algorithms for ranking related pages. When determining the rank score of a web page, all links are considered equally. PageRank is actually very easy. However, when a simple calculation is repeated hundreds of times, the results can become complicated. PageRank is just one factor in determining which results appear first in a Google search. To improve the performance of these methods, a number of algorithms have been developed. The paper [1] introduces the IPR algorithm, which is a PageRank extension. IPR distributes rank scores based on the reachability of the pages, taking into account the respective domains of both the in-links and out-links.

The standard version of PageRank is well established algorithm for indexing of web-pages, however the improved version only works on a particular keyword. It only suggests how removing mutual reinforcement can improve the PageRank algorithm.

VII.ADVANTAGES

i. It considers reachability factor while calculating the rank.
ii. It reduces mutual reinforcement.
iii. It does not consider redundant hyperlinks.
iv. It gives more relevant page for the users' search.

VIII.DISADVANTAGES

i. It has only considered one keyword.
ii. It is not well established.
iii. PageRank has no way of knowing which of these main phrases is the most relevant to the user.
iv. If your website contains circle references, it will lower the PageRank of your front page.

IX.CONCLUSION

This improved version gives better results than standard PageRank algorithm for a keyword that is looked up for in a domain. In the current version of IPR, the rank score was calculated using only different domain in-links, out-links, and reachability. However, this model has a drawback in that if a web site owner creates a connection to a different domain from the same domain or to the same web page, this method would give that webpage a higher ranking.

DOI 10.17148/IJARCCE.2021.105142

We can conduct a comprehensive performance analysis of IPR using other keywords in the future, as well as categorise web pages by increasing the number of 'human' users. Furthermore, a thorough study of IPR output using various keywords is possible.

REFERENCES

[1]. R. Sardhara, K. I. lakhataria, An Improved PageRank Algorithm Based on Reachability to Reduce Mutual Reinforcement Effect , IEEE, 10th ICCCNT 2019 July 6-8, 2019, Kanpur, India.

[2]. Brin, S. and L. Page, "The anatomy of a large-scale hypertextual Web search engine". Comput. Netw.ISDN Syst., 1998. 30(1-7): p. 107-117.

[3]. Ian Rogers, The Google PageRank Algorithm and How It Works, IPR Computing Ltd.

[4]. Miguel Gomes da Costa Júnior and Zhiguo Gong, "Web Structure Mining: An Introduction " , Proceedings of the 2005 IEEE, pp 590 – 595.

[5]. Shailendra G. pawar and Pratixa Natani, " Effective utilization of Page Ranking and HITS in Significant Information Retrieval " , International Conference for Convergence of Technology -2014,pp 1 – 6.

[6]. Ashish Jain, Rajeev Sharma, Gireesh Dixit and varsha Tomar, "Page Ranking Algorithms in Web Mining, Limitations of Existing methods and a New Method for Indexing Web Pages" International Conference on Communication Systems and Network Technologies ,2013 ,pp 640 – 645.

[7]. Anurag kumar,Ravi kumar singh "A study on Web Structure Mining " (IRJET) International Research Journal of Engineering and Technology, Vol 04 Issue 1 pp 715-720