International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 7 Issue 6, November-December 2023 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

Data Processing in Web Mining Structure by Hyperlinks and Pagerank

Ku Nalesh[1], Ghanshyam Sahu[2], Lalit Kumar P Bhaiya[3]

[1]M Tech Scholar (CSE), [2]Department of Computer Science & Engineering,
[1,2]Bharti College of Engineering and Technology, Durg, Chhattisgarh, India
[3]Bharti University, Durg, Chhattisgarh, India

ABSTRACT

Creating a quick and effective page ranking system for web crawling and retrieval is still a difficult problem. We suggest constructing a set of PageRank vectors biased using a collection of representative subjects in order to better capture the idea of relevance with regard to a certainty of topic in order to produce more accurate for search results. The outcome of the experiment demonstrates that the suggested algorithm improves the degree of relevance compared to the original one and reduces the topic sensitive PageRank's query time efforts. This paper offers an overview of Web mining as well as a review of its various categories. Next, we concentrate on one of these subcategories: Web structure mining. In this area, we describe link mining and examine PageRank, two well-liked techniques used in web structure mining.

KEYWORDS: *Web Mining, Web Structure Mining, Web Graph, PageRank, Data Processing*

How to cite this paper: Ku Nalesh | Ghanshyam Sahu | Lalit Kumar P Bhaiya "Data Processing in Web Mining Structure by Hyperlinks and Pagerank" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-7 | Issue-6, December 2023, pp.223-228, URL: www.ijtsrd.com/papers/ijtsrd60083.pdf

IJTSRD60083

I. INTRODUCTION

Web structure describes how a website's pages, navigation, and content are arranged and laid out. It includes the connections between the different web sites and how visitors can navigate between them[1]. For a website to offer a great user experience and to guarantee that visitors can quickly locate the information they need, it must be well-structured. Web mining can be thought of as general data mining methods applied to the Web. However, we must significantly adapt and extend the conventional approaches due to the inherent characteristics of the Web. First off, despite the vast amount of data on the Web, it is dispersed online. We must collect the Web document first before mining. Second, because Web sites are semi-structured, documents need to be retrieved and displayed in some way in order to facilitate processing[1, paper 10]. Thirdly, since the meaning of information on the Web might vary, a training or testing data collection should be sufficiently large. Despite the aforementioned challenges, there are still alternative ways to encourage mining on the Web. For instance, it is important to exploit the linkages between Web pages.

A global information system, World Wide Web is distributed by multiple Web sites worldwide. The quantity of web pages is particularly hard to count because web servers have the capacity to hold millions of pages[2]. Web networks resemble tens of thousands of interconnected, entangled cells arranged in a complicated configuration. Each web page on a website has the following three components: a body, hypertext markup, and links to other web pages. Web pages and general documents are different from the material they contain in each web site. Web pages on a website do not exist in a vacuum; they are typically connected through hyperlinks between documents[3].

A website's upkeep is equally as crucial as its construction. We must enhance the website's design in order to retain it. By examining user browsing patterns, we may discover out how a website is used and make necessary design changes.

Users' web browsing habits can be examined using statistical analysis and web usage mining. Page Views, Page Browsing Time, and soon are included in the statistical analysis's conclusion. Web usage mining uses data mining techniques to analyse web usage data to find patterns in web usage. Data mining techniques that can be used to examine online usage data include Item-Set Mining , Sequential Pattern Mining and Graph Mining[4].

Lets say, The URL route to the directory structure of the string of information; Web content can be utilised within the HTML, XML's tree structure represented; and the hyperlink structure between web pages are among the information contained in web structure. A significant amount of web content buried away on the web can be found by mining structural data. Link to main page for the content of Web pages with numerous pages in the navigation to perform the function of Web pages by analysing the internal tree structure, the structure can be obtained and used to find the page with the specified set and 1 2,...., P P Pn content-related pages.

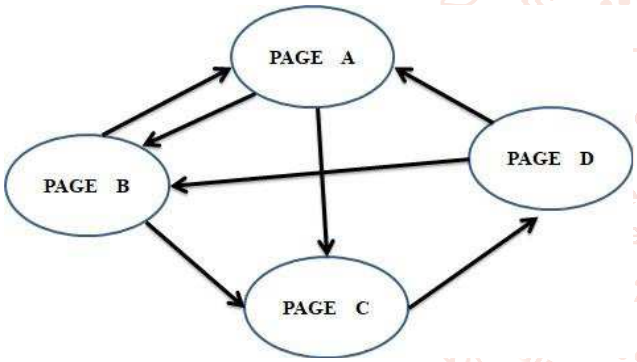


Figure 1 A simple web link graph

Web structure describes how a website's pages, navigation, and content are arranged and laid out. It includes the connections between the different web sites and how visitors can navigate between them[5]. For a website to offer a great user experience and to guarantee that visitors can quickly locate the information they need, it must be well-structured. Web mining can be thought of as general data mining

Lets say, these paper main focus of hyperlink analysis is web structure mining, which involves looking at how pages are connected to one another to discover patterns that can be used to enhance search results[6]. Structure mining refers to the site that has links from one page to another in a link diagram, whereas content mining refers to the site that has every text character, word, and image in the index. Structure mining information included in the file, regardless of content, does not require knowledge of the link between the files.

II. Related Work

A. Overview of Web Structure Mining

To unambiguously define the various sorts of web mining. The following tasks were offered as a dissection of web mining by Kosala and Blockeel [7]:

1. **Resource discovery**: the process of locating desired Web documents.
2. **Information selection and Pre processing :** Automatically picking out and pre-processing particular information from retrieved Web resources.
3. **Generalisation:** automatically identifies broad patterns on both a site-by-site and across-site basis.
4. **Analysis:** verification or interpretation of the patterns discovered during mining.
5. **Data Collection :** Gathering the necessary data for analysis is the initial stage in every mining technique. Collecting data for web structure mining involves gathering website linkages.
6. **Preprocessing:** This is one of the data mining strategies used to make the data as structured and pertinent to the needs as possible.
7. **Knowledge discovery :** It is the process of using numerous data mining techniques, such as statistical explanation, association, clustering, pattern analysis, and similar ones, to the processing of data.
8. **Knowledge analysis :** It is used to select relevant data from the internet and provide information that is more in line with user needs.

B. Connections Between Hyperlink and Page Content

There will be a subject that is more than just a focal point; the World Wide Web sites or pages will not all revolve around the same core (Hub/Central) phase. The link between the most recent close association and the content that is most similar to that content will eventually fade from the gathering centre site or page [8]. Additionally, when the number of linkages is decreased, the association's content gradually shifts from one theme to another. A website is considered a hub if it links to several other authoritative websites, whereas an authoritative website is one that is used by many other hub websites.

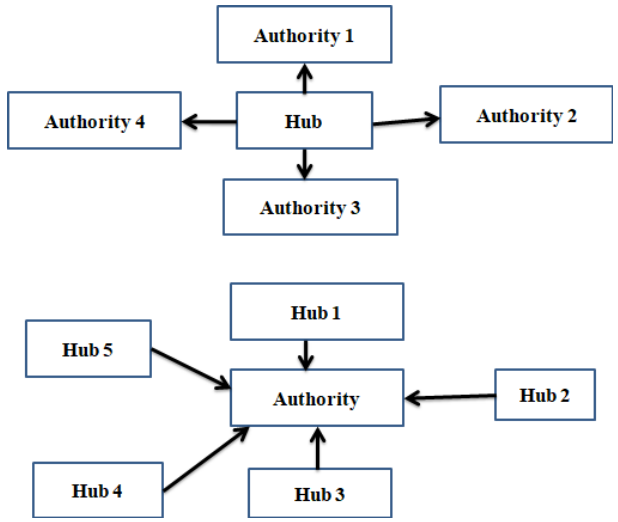


Figure 2 Relationship between Hyperlink and Page Content

Scores are assigned to both hubs and authorities. An authority that is referred to by multiple highly rated hubs should be a powerful authority, and a hub that points to several highly rated authorities should be a well-liked hub [9]. Let ap and hp stand in for the page p authority and hub scores, respectively. The set of referrer and reference pages for page p are indicated, respectively, by B(p) and I(p). The following formulas are used to calculate hub and authority scores [10].

$$a_p = \sum_{q \in B(p)} h_q$$

$$h_p = \sum_{q \in I(p)} a_q$$

C. Page Rank Algorithm

The primary method for analysing web hyperlink structures uses the creation of the Web as a directed graph in accordance with some heuristic criteria and graph theory. The linking structure of one of the most effective representatives is the Page Rank algorithm. Google's search engine uses an algorithm, anchor text tags, word frequency, and other criteria to combine a huge number of search results and determine which ones are most relevant and should appear first [11].

Page Rank, asserts that if a website has important links pointing to it, then those links also become essential for other pages that the page links to. A page has a high rank if the total of the ranks of its backlinks is high [12], and Page Rank considers backlinks when determining ranking through links. Figure 3 displays an illustration of a backlink: Page A links back to pages B and C, whereas pages B and C link back to page D.

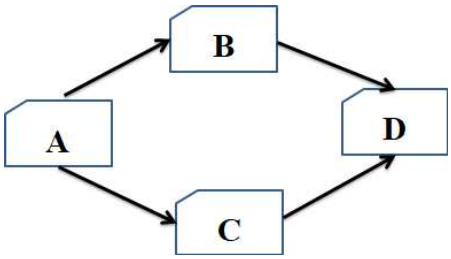


Figure 3 Sample of Back links

$$PR(j) = (1-d) + d\sum_{i \in B_i} \frac{PR(i)}{|F_i|}$$

III. Proposed Method

We looked at the user's behaviour to find a solution to the rank sink issue. It has been observed that not all users click on the links that are already there. For instance, after seeing page a, some people might choose not to click on the links that are already there and instead proceed directly to page b, which is not connected to page a. Users can accomplish this by entering page b's URL into the URL text field, which will take them directly to page b. Although these two pages are not directly related in this instance, page should nevertheless have an impact on page b's ranking. Consequently, there isn't an absolute rank sink.

$$PR(u) = (1-d) + d \sum_{v \in B(u)} \frac{PR(v)}{N_v}$$

Taking into account the phenomenon, where d is a dampening factor that is often set at 0.85. We might also use (1 d) to be the pagerank distribution from non-directly linked pages and think of d as the likelihood of visitors clicking on the links [13].

Google applied the PageRank algorithm to the Google search engine to evaluate its usefulness. Because the rank value converges to a suitable tolerance in the nearly logarithmic (log n) [space], the PageRank method functions well and efficiently in the trials.

A web page's rank score is distributed equally across the pages it links to. Even while Google successfully uses the PageRank algorithm, there is one issue: on the actual web, some links on a web page may be more essential than are the others.

IV. Experiments and Result

Page Rank is implemented as follows: after pretreatment, each page's initial PR value is set to 0, and through the aforementioned recursive algorithm, each page's Page Rank value is repeated iteratively until the results converge. This process converts each hyperlink's integer ID to the index's integer ID and corresponds to the page's URL. This vector is

independent of the search query and is computed once offline. Without taking into account the relative value of the various connections, the PageRank algorithm gives external links an average weight for their contribution.

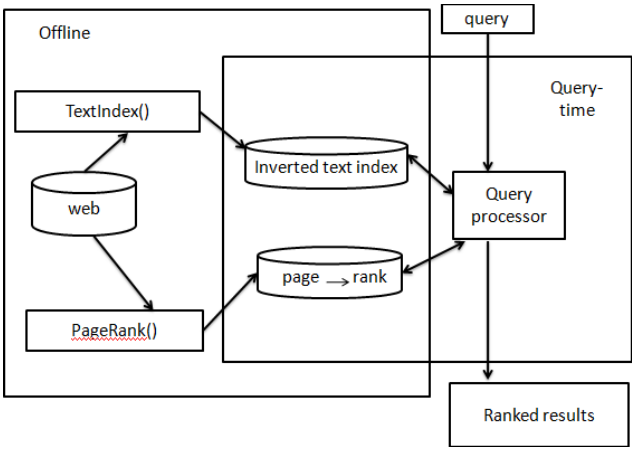


Figure 4 PageRank External Working with Query

Step 1: Hub weight is indicated on page p of the Authority document as T Fu initial all pages.

ap ←1, hp←1 initialize

Step 2: The iterative formula below and continued adjustments until the outcomes are consistent:

$$a_p = \sum_{\forall q:q\to p} h_q$$

$$h_p = \sum_{\forall q:q\to p} a_q$$

The implication of the symbol p←q is that the point p from the q hyperlink exists. The end output of algorithm is a set of weights with large Hub p pages and have greater weight Authority page after multiple repeated calculations until the results converge[14].

Web pages are viewed as a vast directed diagram by the PageRank algorithm, which only takes into account links between pages.

In this paper we describes graphs of Web structure mining, which looks at how pages are connected to one another to find patterns that can be utilised to improve search results, is the main emphasis of hyperlink analysis.In contrast to content mining, which refers to websites with every text character, word, and image in the index, structure mining refers to websites with linkages connecting pages in a link diagram. Regardless of its content, structure mining information in a file does not need to be aware of how the files are connected.

{
"Website A": ["Website B", "Website C"],
"Website B": ["Website C", "Website D"],
"Website C": ["Website D", "Website E"],
"Website D": ["Website E"],
"Website E": []

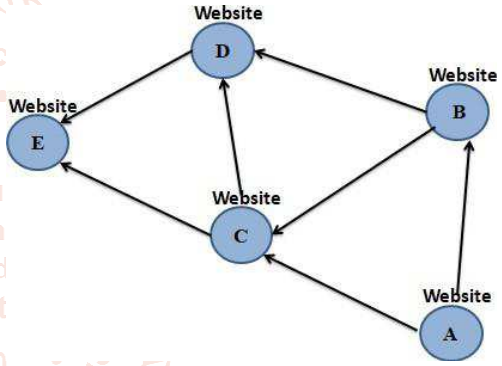
Website A: 0.09186427056562586
Website B: 0.13090643205037966
Website C: 0.18654161663019636
Website D: 0.2267797942548725
Website E: 0.36390788649892575

PageRank scores : 0.1999

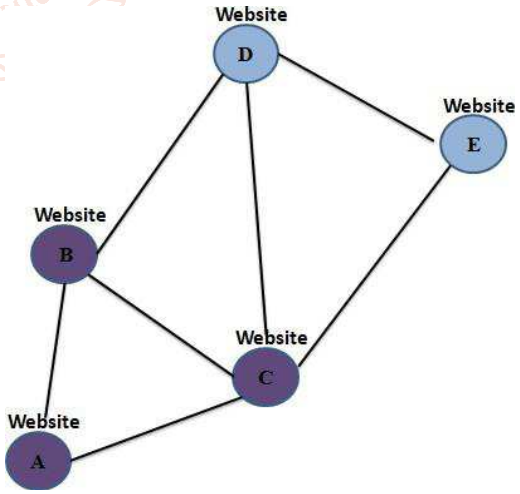
Graph Result

We deployed conventional PageRank algorithms to compare the outcomes in order to assess the PR method. The various parts that go into implementing and assessing the PR algorithm are shown in below Figures **5**.

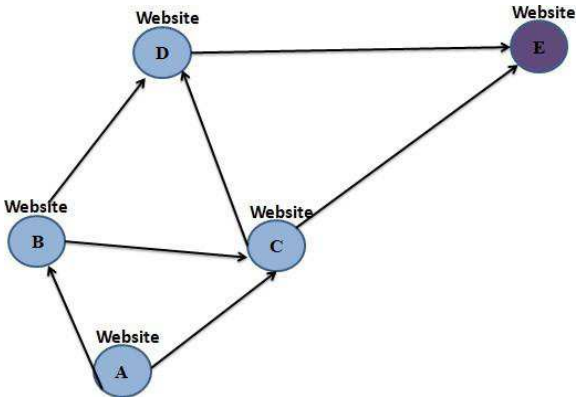
Graphs



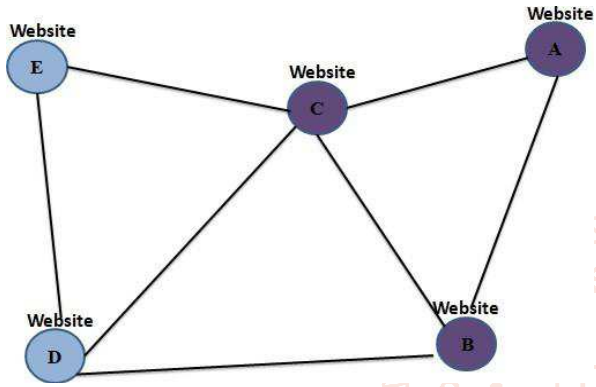
Graph 1 Sample Web graph Visualization with Label Propagation Communities



Graph 2 Web graph with Louvain Modularity



Graph 3 Web graph Visualization with PageRank Scores



Graph 4 Web graph Visualization with Louvain Modularity(Modularity Score : 0.0306)

The six main actions that make up the simulation studies we conducted for this work are as follows:

1. **Locating a website:** Because the typical PageRank algorithms depend on the web structure, locating a website with rich hyperlinks is important.
2. **Creating a web map:** This website does not have a web map. To create the necessary web map, JSpider, a free spider software, is employed.
3. **Identifying the root set**: Using the IR search engine built within the website, a collection of pages pertinent to a specific query are retrieved. The root set is this collection of pages.
4. **Locating the base set:** A base set is produced by enlarging the root set with additional pages that either directly point to or are pointed to by the root set pages.
5. **Using algorithms:** The WPR and Standard PageRank algorithms are used on the initial set.
6. **Analysing the outcomes:** The algorithms are assessed by contrasting their outcomes.

V. Conclusion:

In this paper, we'll use an PaperRank algorithm to preprocess the data. All of the necessary urls will have their hyperlinks extracted by this technique. The algorithm preprocesses the information from these hyperlinks, removing any material that is unnecessary for the user and providing the user with the pertinent data they need. By eliminating the useless material, the final result is to provide the user with useful information. The information era has begun as a result of the exponential growth of web information. In this work, one of the most successful Page Rank algorithms is enhanced for distinct hyperlinks allocated varying weights in order to improve the current algorithm, which now assigns different hyperlinks the same weight. However, even with better Page Rank, web page analysis is still not dependent on the content of web pages. Future study will concentrate on how Web structure mining weights web text for the associated hyperlink based on relevancy. In addition, there are other issues that merit additional study, such as how to determine the values of the damping factors discussed in the previous sections and how to create a functional expression for the relationship between similarity and divergence.

References:

[1] Fahd Alhaidari, Sarah Alwarthan, Abrar Alamoudi, "User Preference Based Weighted Page Ranking Algorithm", IEEE, 2020

[2] Rasha Hani Salmana Mahmood Zakib Nadia A. Shiltag c, "A STUDYING OF WEB CONTENT MINING TOOLS", Al-Qadisiyah Journal of Pure Science Vol.(25) Issue (2) pp. Comp. 1–16,2020

[3] Ayyavaraiah Monelli, Shoban Babu Sriramoju, "An Overview of the Challenges and Applications towards Web Mining", Second International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC 2018) IEEE Xplore, ; ISBN:978-1-5386-1442-6, 2018

[4] Dheeraj Malhotra & O. P. Rishi, "A comprehensive review from hyperlink to intelligent technologies based personalized search systems", Journal of Management Analytics, ISSN: 2327-0012, 2019

[5] Bin Liu, Shuangyan Jiang and Quan Zou, "HITS-PR-HHblits: protein remote homology detection by combining PageRank and Hyperlink-Induced Topic Search", Briefings in Bioinformatics, Published by Oxford University Press,2020.

[6] Muhammd Jawad Hamid Mughal, "Data Mining: Web Data Mining Techniques, Tools and Algorithms: An Overview", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 9, No. 6, 2018

[7] Atsushi Suzuki and Hideaki Ishii , "Distributed Randomized Algorithms for PageRank Based on a Novel Interpretation", Annual American Control Conference (ACC) June 27–29, 2018. Wisconsin Center, Milwaukee, USA,2018

[8] Prem Sagar Sharma , Divakar Yadav and R. N. Thakur, "Web Page Ranking Using Web Mining Techniques: A Comprehensive Survey", Hindawi Mobile Information Systems Volume 2022,2022

[9] Victoria Kaysera, Erduana Shalaa, "Scenario development using web mining for outlining technology futures", Technological Forecasting & Social Change, Elsevier 2020

[10] Raghavendra , K G Mohan, "Web Mining and Minimization Framework Design on Sentimental Analysis for Social Tweets Using Machine Learning", International Conference on Pervasive Computing Advances and Applications, Procedia Computer Science 152 (2019) 230–235, Elsevier,2019

[11] Hui Li, "Internet Tourism Resource Retrieval Using PageRank Search Ranking Algorithm", Wiley Hindawi Complexity Volume 2021, Article ID 5114802, 11 pages, 2021

[12] Kareem K. Ibrahim - 2 Ahmed J. Obaid, "Web Mining Techniques and Technologies: A Landscape View", Ibn Al-Haitham International Conference for Pure and Applied Sciences, Journal of Physics: Conference Series 1,2021

[13] Shiqi Zhang, Renchi Yang, Xiaokui Xiao, Xiao Yan and Bo Tang, "Effective and Efficient PageRank-based Positioning for Graph Visualization", Proc. ACM Manag. Data, Vol. 1, No. 1, Article 76. Publication date: May 2023.

[14] Leandro Tortosaa , Jose F. Vicent a,∗ , Gevorg Yeghikyanb, "An algorithm for ranking the nodes of multiplex networks with data based on the PageRank concept", Applied Mathematics and Computation, 392 (2021) 125676, Elsevier, 2021