DOI : https://dx.doi.org/10.26808/rs.ed.i10v2.03
International Journal of Emerging Trends in Engineering and Development Issue 10, Vol.2 (Feb-Mar 2020)
Available online on http://www.rspublication.com/ijeted/ijeted_index.htm ISSN 2249-6149

WEB PATTERN ANALYSIS USING PARTITIONING ALGORITHM IN HYPERLINK STRUCTURE

MR. AVINAV PATHAK[#1], Dr. SAURABH SHARMA[#2], Mr. RAJESH PANDEY[#3]
#1Asst.Prof., CSE, Shobhit University, Meerut, 9456089798 and
avinav.pathak@shobhituniversity.ac.in
#2 Assoc. Dean, Vidya Knowledge Park, 7830202777 and saurabhsharma@vidya.edu.in
#3 Asst. Prof., CSE, Shobhit University, Meerut, 7417970699 and
rajesh@shobhituniversity.ac.in

ABSTRACT: This research paper focuses on the current hot topic web pattern analysis which is also useful in digital marketing (Search Engine Optimization (SEO), Social Media Optimization (SMO). It also covers the webometrics algorithm analysis and other ways used to analyze the web pattern using hyper linking structure. A number of ways have also been survey like hyperlinking structure analysis and like image compression/size formatting, minifying scripting of webpage, combining different types of scripts to reduce hyperlinking load. A comparative study between Page Rank, Weighted Pagerank and HITS is also done in order to give an in depth insight to the factors affecting the same. Ultimately, the objective to find out an efficient way to optimize the web pattern analysis process to improve its overall reach. An algorithm has also been proposed for analyzing web pattern using hyperlinking structure.

Key words: Pagerank, Weighted Pagerank, HITS, SEO, SEM, SMO etc.

Corresponding Author: Dr. Saurabh Sharma

1. INTRODUCTION

In view of the rapidly increasing use of the Internet and the amount of data available on the Internet, it has become very important to determine which data is relevant and which is irrelevant. The web search has become incredibly powerful, giving it almost any kind of information within the billions of pages that make up the web, can be tracked down and used. Nowadays almost every search engine faces the increasingly difficult challenge of collecting, storing, processing, retrieving and distributing web data for users with different purposes, needs and search backgrounds. While conventional algorithmic search engines have been very successful in processing web searches with relatively simple keywords, recently due to the emergence of new user groups and user requirements there is a great deal of interest in new generation internal search that requires the development of new web search applications [1]. The two predominant paradigms for finding information on the Web are navigation and search [2]. Most web users often use a web browser to navigate a website. They start with the homepage or a webpage found through a search engine or linked from another website, and then follow the hyperlinks they consider relevant on the homepage and subsequent pages, until they find the desired information on one or more pages. They can also use the search functions provided on the website to speed up the search for information. For a website that consists of a large number of web pages and hyperlinks between them, these methods are not enough for users to find the desired information effectively and efficiently. The two predominant paradigms for finding information on the Web are navigation and search [2],

[3]. Most Web users typically use a Web browser to navigate a website. They start with the home page or a Web page found through a search engine or linked from another Web site, and then follows the hyperlinks they think relevant in the starting page and the subsequent pages, until they have found the desired information in one or more pages. They may also use search facilities provided on the Web site to speed up information searching. For a Web site consisting of a very large number of Web pages and hyperlinks between them, these methods are not sufficient for users to find the desired information effectively and efficiently. The section 2 contains in-depth background study following with the most used webomatrics and analysis of Pagerank, Weighted Pagerank, & HITS. Then the successive sections deals with the comparative study in a tabular form. The proposed algorithm is discussed following with conclusion in later sections.

2. BACKGROUND STUDY

One of the main objectives here is the analysis of connection structure algorithms thus, different algorithms are analyzed in order to accomplish it. The algorithms are Page Rank [13], hypertext-induced topic search (HITS), weighted Page Rank[12] and other variations of Page Rank as shown in table 1 of section 5. This objective also includes the performance analysis of various types of classification algorithms (manual, based on assumptions). We have briefly analyzed all these algorithms. Some of the algorithms are analyzed taking into account graphics with incoming and outgoing links. The data is imaginary and not based on facts. Inbound links, which are also called inbound links, inbound links, back links and inbound links, are inbound links to a website or website. In the basic terminology of the link, an incoming link is a link that receives a web node (website, directory, website or top-level domain) from another web node [14]. Web nodes are web pages of the website. Each website is a collection of websites or nodes. A website consists of a series of websites related to content such as text, images, video, audio, etc. A website is hosted on at least one web server that can be accessed. A network such as the Internet or a private local area network through an Internet address called Uniform Resource Locator [11]. A website is a collection of web pages (documents accessed through the Internet). A web page appears on the screen when you enter a web address, click on a link or search in a search engine. A web page can contain any type of information and can contain text, colors, graphics, animations and sound. When someone gives you your web address, they will usually direct you to the home page of your website, which will tell us what that website offers in terms of information or other services. On the home page, you can click on the links to go to other sections of the site. A website can consist of one page or tens of thousands of pages, depending on what the website owner wants to achieve. A website is a document that is usually written in plain text and interspersed with formatting instructions for the hypertext markup language (HTML, XHTML). A website may contain elements of other websites with appropriate marking anchors. An outbound link is an HTML code on your website that allows visitors to access other websites. These are often simply called links. Each time user click on a link in a website that takes them to a different location, especially from your website, you provide external links. There are also some questions about the value of links. Sometimes people argue that they have a problem because they remove people from their websites. Others, especially in the world of search engine optimization (SEO) and content writing, know that links are an important reciprocal gesture [15]. A number of other researches based on computational intelligence are also going on from the field of networking to mathematical analysis in order to automatically generate the intelligent analysis as research done in [7], [8] and [9]. These researches includes Self organizing Maps, Genetic Algorithm and other powerful algorithms which supports the analysis like adaptive Neuro fuzzy Inference System or ANFIS.

. External links must be distinguished from an internal link. The inline is a hyperlink on the website of another person that refers people to their pages. This essentially means that all external links you provide on your website are external links for another person. A good proportion of links and external links help to improve the profile of your website. In particular, links to your website help improve the ranking of your pages in search engines. However, if you do not provide links to other people on your site, they are unlikely to provide you with links. To facilitate its use, alternative and inserted links are very important for each website. This saves the user time and the user can easily access the relevant content on the World Wide Web [11].

3. DETAILED STUDY: HYPERLINK ANALYSIS METHODOLOGY

Based on the previous discussion, research on hyperlink analysis can be classified according to the dimensions of the knowledge model, metrics, scope of analysis and algorithms. In our observation, these form the basic building blocks for hyperlink analysis. In this section, we will first classify a series of applications listed in the literature based on these dimensions, and then we will propose a general methodology for using hyperlink analysis that meets the purposes of an application.

The methodology for using hyperlink analysis for an application can be described as the following sequence of steps:

1. Analyze the application requirements to determine the type of information needed through hyperlink analysis. For example, the web search application requires that the pages relevant to a user's request be classified by importance. The information model here is a URL classification. In some cases, a process model may be required in addition to the information model.

2. Then, determine the metrics that should be calculated to quantify various aspects of the information model. For Google, for example, the metric is Page Rank, while in the HITS approach it is a center score and authority score. As new hyperlink analysis applications are discovered, new metrics must be developed to meet user's needs.

3. The algorithms to calculate the selected metrics should be selected / designed below. The Google method uses a scrolling algorithm (restricted area) to calculate the Page Rank metrics of the pages that are relevant to a user query as shown in table 1 in section 5. The HITS approach has developed a new algorithm called Hypertext Indexing and Theme Selection (HITS), which uses the algorithm to calculate values typical of large and scattered matrices as its main workhorse. The metrics and hyperlink analysis algorithms to calculate them are closely related, and every time a metric is designed, an algorithm for it should generally be designed.
4. The next step is to determine the scope of the analysis relevant to the application. We can choose between individual page levels, page groups and links, or a complete diagram. A similar analysis can be performed with different areas for different applications.
5. Finally, it must be decided whether hyperlink analysis should be performed alone or in relation to web content and web usage analysis. In this case, the results should be integrated with those of the other analyzes.

We believe that following this approach can help in better leveraging the growing body of techniques and experiences with hyperlink analysis.

DOI : https://dx.doi.org/10.26808/rs.ed.i10v2.03
International Journal of Emerging Trends in Engineering and Development Issue 10, Vol.2 (Feb-Mar 2020)
Available online on http://www.rspublication.com/ijeted/ijeted_index.htm ISSN 2249-6149

4. HYPERLINK ANALYSIS: THE WEBOMETRICS APPROACH

The interest of information science in hyperlinks began in 1996 and was mainly determined by analogies with citations in journal articles. Citations are widely used in the fields of bibliometry and Scientometrics to evaluate the quality of scientific work and follow patterns of scientific communication [16, 17]. The underlying assumptions are that more important or higher quality articles tend to be cited more frequently, and that quotes often indicate that the work in the cited article is based or used by the cited article (Cronin, 1984). In fact, the reasons for the appointment are very varied (Borgman and Furner, 2002), but dating analysis remains an effective, although controversial, tool used for a variety of purposes (Garfield, 1979; Oppenheim, 1997;Moed). Two first articles examined the use of hyperlinks to track web information. An extensive theoretical discussion by [18] also laid the groundwork for the new field of webometry and gave it a name. The event that triggered Webometric was the provision by commercial search engines such as AltaVista of an interface that allowed anyone to count links between large web spaces with a simple command. This allowed us to think about techniques to use this new installation and speculate and investigate possible new applications. Information scientists who recognized this potential naturally resorted to their own disciplines to find applications, and the seemingly close analogy between hyperlinks and citations, which refer to one document by another, gave them a set of research questions and techniques. through techniques adapting the analysis of citations.
Then, the starting point of Webometric was the attempt to apply the analysis of citations to the web context. Because citation analysis tracks (to some extent) scientific communication, some researchers have attempted to use hyperlink counts as a measure of the amount of online communication between owners of two or more sets of websites. In other dating analyzes, attempts are made to evaluate work areas based on the number of appointments. This has led to a second type of webometry approach: verifying whether the number of links can be a valid measure of the impact online. This has led to investigations on whether pages attract hyperlinks primarily due to their quality or interest in the content, so that the amount of hyperlinks would
measure some kind of impact online.The methodological starting points are not only formulas and algorithms for the calculation of useful information, but also a wide range of data validation techniques, partly a legacy of the ongoing controversy over the analysis of evaluative citations.Data collection: web crawlers and commercial search engines. Early webometric studies used advanced search engine queries to determine the amount of hyperlinks, mainly AltaVista and AllTheWeb. For example, if you enter the query host: wlv.ac.uk and the link: knaw.nl in the AltaVista advanced query section, you can count the number of pages with domain names ending in wlv.ac.uk and that contain a hyperlink text knaw.nl. This is expected to result in a series of pages from WolverHampton University (http://www.wlv.ac.uk) that provide a hyperlink to the Koninklijke Nederland Academy of Arts and Sciences (KNAW, Royal Dutch Academy of Arts and Science), http: // www.knaw.nl), either the main page or a subdomain. Such consultations allowed easy access to useful data from the huge search engine databases. One of the known drawbacks is that no search engine can index the entire web (Thelwall, 2002) and the actual coverage in 1999 for the main search engines of the time was less than 16% (Lawrence & Giles, 1999). If a commercial search engine is used, this is clearly a limitation that must be accepted and discussed in the study.However, a second important problem was immediately discovered: the results of the search engines fluctuated irregularly, sometimes dramatically (Mettrop and Nieuwenhuysen, 2001; Rousseau, 1999; Snyder and Rosenbaum, 1999; Thelwall, 1999) . To counter this, Rousseau (1999) proposed several search rounds and an average process. In addition, search engines have shown peculiarities in behavior in the past. Snyder and Rosenbaum (1999) reported that AltaVista had a number of problems discussed in coming

section.Through the AltaVista metatermatic link, you can call the total number of pages, each of which contains at least one link to a specific page. In practice, the term goal often cannot access all or even most links. On the contrary, the "link" command sometimes retrieves pages that do not contain the link specified (GOOGLE). The following 3 points contains performance analysis and suggests ways to improve the performance.

4.1 PERFORMANCE ANALYSIS OF WEBSITES

The website is a collection of one or more websites grouped under the same domain name. A review of website performance begins with a review of key indicators. The key indicators provide an overview of the overall performance of the website and an overview of the areas that need improvement [19]. If your current tool indicates that 60% of website visitors leave your instructional video, you may assume that the content is not relevant to your target audience. However, this assumption is incorrect if you are not sure how the interactive elements of your website work. For example, the problem could be with a content delivery network (CDN) in a specific geographic location. To get a complete picture of the functionality and merchantability of your website, you must perform root cause analysis and correlation with both metrics.Website performance is an important issue for many people in an organization. Most companies use some type of tool to monitor the interaction of visitors with their website. Marketing specialists want to know what triggers the most potential customers a blog, a video, a download library, etc. Sales representatives want to receive notification when a potential customer visits the website. It is a common practice to use a series of different metrics to evaluate website performance from a marketing perspective. Unfortunately, most tools do not tell you why a website works better or worse from the perspective of the end user. This is important information for marketing and sales, as well as for a company's IT department.

.

If your current tool reports that 60% of site visitors are abandoning your 'how-to' video, you might assume the content is not relevant for your audience. That assumption is flawed, however, if you are not also acutely aware of how the interactive elements of your site are performing. For instance, the problem could be with a Content Delivery Network (CDN) in a particular geographic location. In order to have a comprehensive picture of both the functionality and the marketability of your site, you need root cause analysis and correlation with both sets of metrics.

To begin, let's first explore traditional metrics.

- **Traditional Metrics**

The following Key Performance Indicators (KPI) for actionable insight used to typically monitor the online presence by leading IT companies :

1. Value per Visitor (Sales / Total Number of Visitors)
2. Conversion Rate (Desired Action / Total Number of Visitors)
3. Cost per Lead (Money Spent / Total Number of Leads Produced
4. Cost per Visitor (Money Spent / Total Number of Visitors)
5. Cost per Customer (Cost per visitor X the number of unique visitors needed to produce a sale)

While these performance metrics [20] are important and quite useful, there could be any number of underlying reasons that they are skewed and thus lead companies to make misinformed business decisions. To be truly useful, companies need to look at the whole picture and look for root causes, not aggregate results.

DOI : https://dx.doi.org/10.26808/rs.ed.i10v2.03
International Journal of Emerging Trends in Engineering and Development Issue 10, Vol.2 (Feb-Mar 2020)
Available online on http://www.rspublication.com/ijeted/ijeted_index.htm ISSN 2249-6149

- **Root Cause Analysis - End-User Experience Metrics**

In most cases, several departments need to work together. For example, while marketing concludes that one promotion is more effective than another, IT can help identify technology problems (or solutions) that affect the results of the promotion. If a cart payment transaction is canceled immediately before the credit card application, the price may be the cause. However, if transactions on certain product pages are systematically interrupted, it may actually be the product, marketing, advertising or technology that caused the interruption. Do you know why users abandoned your websites? Was it because of a marketing error or a technical problem? The main factor is the measurement of the USER's experience. This requires new tools that see things from the perspective of the end user. From the perspective of the browser. Among other things, you should evaluate how fast a page loads from a visual perspective and how fast a page loads for the application to be available.
It has also been seen affected and isolate performance bottlenecks by evaluating the underlying impacts from:

1. Application Performance
2.3rd party content load (and play) performance
3. Network latency (e.g. testing different regions of the country and evaluating CDN performance)
4.Database Performance
5. Infrastructure Performance - down to the individual element levels

Following online parameters are mandatorily affects the performance and may be used as the main variable to check typical online goals, it includes: Page views Number of Downloads (PDFs, Whitepapers, Applications, Multimedia files, Flash files, Blog feed, etc.), Forms (Search boxes, Subscribe forms, Contact form, etc.) , Purchase and Time Spent on the website.
To get the website analysis, we need web analytics tools which have following capabilities:

1. Which search engines are referring people to the website and the keywords that they use?
2. How many times are the pages viewed?
3. What pages do visitors frequent?
4. How long do these visitors spend on the website?

Few ways are given below which played a role of game changer in web pattern and performance analysis.

4.2 Ways to Improve Website Performance

There are a million and one way to improve the performance of your website. The methods vary and some are more complicated than others. The three main areas in which you can work are: hardware (your web server), server-side script optimization (PHP, Python, Java) and front-end performance (the meat of the website). (for example, HTML, CSS, JavaScript and images) is the most accessible part of your website. If you use a shared web hosting plan, you may not have root (or similar to root) access to the server and therefore cannot optimize and adjust the server configuration. And even if you have the correct permissions, the development of web servers and databases requires specialized knowledge to obtain immediate benefits. If the website speed is good and works correctly, it will be classified as not. of users. It is easy to browse the web if the structure is not complex. Below are some simple ways to improve the speed of your website as shown in Figure 1 below.

DOI : https://dx.doi.org/10.26808/rs.ed.i10v2.03
International Journal of Emerging Trends in Engineering and Development Issue 10, Vol.2 (Feb-Mar 2020)
Available online on http://www.rspublication.com/ijeted/ijeted_index.htm ISSN 2249-6149

- **Profile your web pages to find the culprits**

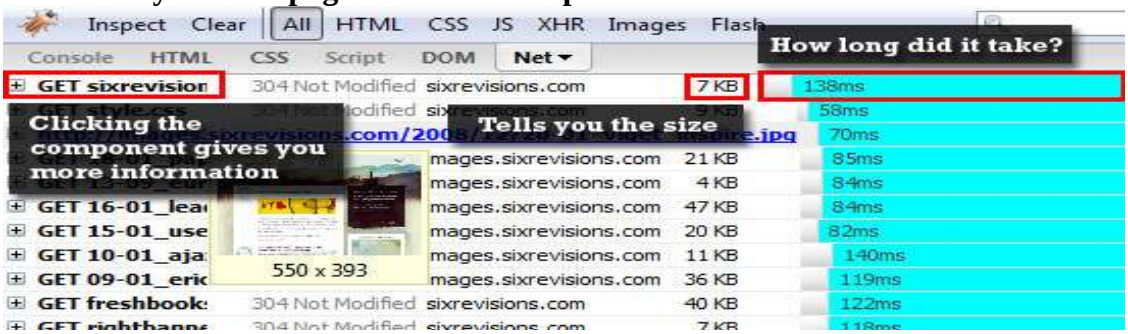

Figure 1. WebPages culprit

It is useful to create a profile of your website to find components that you do not need or that can be optimized. Profiling a website generally requires a tool like Firebug to determine what components (i.e. images, CSS files, HTML documents and JavaScript files) are requested by the user, how long it takes for the component to load and how large it is. As a general rule, keep the components of your page as small as possible (less than 25 KB is a good goal) as shown in fig1.
The Firebug Network tab (see above) can help you find large files that can damage your website. In the example above, you can see that you get a breakdown of all the necessary components to represent a website, including: what it is, where it is, how big it is and how long it took to load
.

- **Save images in the right format to reduce their file size**

Figure 2. Format for image saving

If you have many images, be sure to determine the optimal format for each image. There are three popular web image file formats: JPEG, GIF and PNG. In general, you should use JPEG for realistic photos with smooth gradients and tones. You must use GIF or PNG for monochrome images (for example, diagrams and logos). GIF and PNG are similar, but PNG generally produces a smaller file size. Read the horror coding weighing with PNG over/on GIF as shown in figure 2 above.

- **Minify your CSS and JavaScript documents to save a few bytes.**

Minification is the process of removing unneeded characters (such as tabs, spaces, source code comments) from the source code to reduce its file size. For example:
This chuck of CSS:

```
.some-class {
  color: #ffffff;
  line-height: 20px;
  font-size: 9px;
```

```
}
```

can be converted to:

```
.some-class{color:#fff;line-height:20px;font-size:9px;}
```

…and it'll work just fine.

And don't worry; you don't have to manually reformat your code. A variety of free tools are available to minimize your CSS and JavaScript files. For CSS, you can find a series of easy-to-use tools in this list of CSS optimization tools. For JavaScript, some popular minimization options are JSMIN, YUI Compressor and JavaScript Code Improver. With a good minimization application, you can undo the minimization when it is under development. Alternatively, you can use a browser tool like Firebug to display the formatted version of your code.

- **Combine CSS and JavaScript files to reduce http requests**

An HTTP request to the server is created for each component that is required to represent a website. So, if you have five CSS files for a website, you need at least five separate HTTP GET requests for that particular website. The combination of files reduces the amount of HTTP requests needed to generate a web page. Read Niels Leenheer's article on how to combine CSS and JS files with PHP (which can be adapted to other languages). Site Point describes a similar method of JavaScript. They saved 1.6 seconds of response time, reducing the response time by 76% of the original time.
Otherwise, you can combine your CSS and JavaScript files using good, old copy-and-pasting.

5. PAGERANK, WEIGHTED PAGERANK AND HITS: ACOMPARISION

5.1 Pagerank

The Pagerank algorithm , weighted Pagerank & HITS are being differentiated on the basis of many parameters like ***Mining Technique Methodology Input Parameter Relevancy Quality of Results, Importance and Limitation*** as shown in table 1. But we must concentrate first to understand them individually. Web may be considered to resemble as a graph as assumed in the following figure:

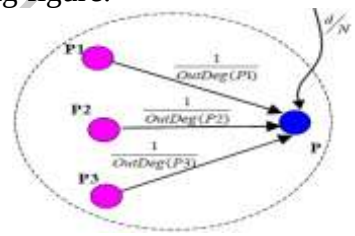

Figure 3 Graph for showing PageRank

In previous study, L. Page and S. Brin [21] proposed the Page Rank algorithm to calculate the importance of web pages using the link structure of the web. In their approach Brin and Page extends the idea of simply counting in-links equally, by normalizing by the number of links on a page. The Page Rank algorithm is defined as: "We assume page A has pages P1...Pn 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. C (A) is defined as the number of links going out of page A. The Page Rank of a page A is given as follows:

PR(A)=(1-d)+d(PR(P1)/C(P1)+...PR(Pn)/C(Pn))

DOI : https://dx.doi.org/10.26808/rs.ed.i10v2.03
International Journal of Emerging Trends in Engineering and Development Issue 10, Vol.2 (Feb-Mar 2020)
Available online on http://www.rspublication.com/ijeted/ijeted_index.htm ISSN 2249-6149

Note that the Page Ranks form a probability distribution over web pages, so the sum of all web pages' Page Ranks will be one." And "The d damping factor is the probability at each page the "random surfer" will get bored and request another random page."

5.2 Weighted Pagerank

Weighted PageRank Algorithm is proposed by Wenpu Xing and Ali Ghorbani. Weighted page rank algorithm (WPR) is the modification of the original page rank algorithm. WPR decides the rank score based on the popularity of the pages by taking into consideration the importance of both the in-links and out-links of the pages.
Weighted Page Rank assigns large rank value to more important pages instead of dividing the rank value of a page evenly among its outlink pages[14].

The popularity from the number of inlinks and outlinks is recorded as $W^{in}(v,u)$ and $W^{out}(v,u)$, respectively.

$W^{in}(v,u)$ is the weight of *link*(*v*, *u*) calculated based on the number of inlinks of page *u* and the number of inlinks of all reference pages of page *v*.

$$W^{in}(v,u) = (Iu) / \sum_{p \epsilon R(v)} Ip \quad \ldots\ldots\ldots (2)$$

Where *Iu* and *Ip* represent the number of inlink of page *u* and page *p*, respectively.
R (*v*) denotes the reference page list of page *v*.

$W^{out}(v,u)$ is the weight of *link*(*v*, *u*) calculated based on the number of outlink of page *u* and the number of outlink of all reference pages of page *v*.

$$W^{out}(v,u) = Ou / \sum_{p \epsilon R(v)} Ip \quad \ldots\ldots\ldots (3)$$

Where *Ou* and *Op* represent the number of outlinks of page *u* and page *p*, respectively. *R*(*v*) denotes the reference page list of page *v*.
Considering the importance of pages, the original PageRank formula is modified as

$$PR(u) = (1-d) + d \sum PR(v) W^{in}(v,u) W^{out}(v,u) \quad \ldots.(4)$$

5.3 HITS

HITS algorithm ranks the web page by processing in links and out links of the web pages. In this algorithm a web page is named as authority if the web page is pointed by many hyper links and a web page is named as HUB if the page point to various hyperlinks. An Illustration of HUB and authority are shown in figure 4

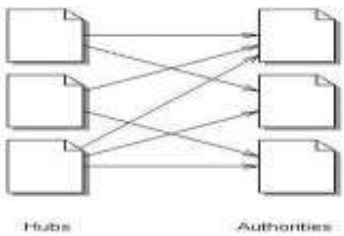

Figure 4 Hubs and Authorities

In HITS [9] algorithm, ranking of the web page is decided by analyzing their textual contents against a given query. A page might be a good hub and a good authority in the same time. This circular relationship leads to the definition of an iterative algorithm, HITS.

DOI : https://dx.doi.org/10.26808/rs.ed.i10v2.03
International Journal of Emerging Trends in Engineering and Development Issue 10, Vol.2 (Feb-Mar 2020)
Available online on http://www.rspublication.com/ijeted/ijeted_index.htm ISSN 2249-6149

Following expressions are used to calculate the weight of Hub *(Hp)* and the weight of Authority *(Ap)*.

$$Hp = \sum_{q \epsilon I(p)} Aq \qquad Ap = \sum_{q \epsilon B(p)} Hq$$

Table 1 Comaparitive Study Of Link Analysis Algorithms

ALGORITHM-> / CRITERIA	**PageRank**	**HITS**	**Weighted PageRank**
Mining Technique	WSM	WSM & WCM	WSM
Methodology	This algorithm computes the score for pages at the time of indexing of the pages.	It computes the hubs and authority of the relevant pages. It relevant as well as important page as the result.	Weight of web page is calculated on the basis of input and outgoing links and on the basis of weight the importance of page is decided.
Input Parameter	Back links	Content, Back and Forward links	Back links and Forward links.
Relevancy	Less (this algorithm Rank the pages on the indexing time)	More (this algorithm Uses the hyperlinks so according to Henzinger, 2001 it will give good results and also consider the Content of the page)	Less as ranking is based on the calculation of Weight of the web page at the time of indexing.
Quality of Results	Medium	Less than PR	Higher than PR
Importance	High. Back links are considered.	Moderate. Hub & authorities scores are Utilized.	High. The pages are sorted according to the importance.
Limitation	Results come at the time of indexing and not at the query time.(Query Independent)	Topic drift and efficiency problem	Relevancy is ignored.(Query Independent)

Here Hq is Hub Score of a page,
Aq is authority score of a page,
I(p) is set of reference pages of page p and *B(p)* is set of referrer pages of page p.
The main drawback of this algorithm is that the hubs and authority score must be computed iteratively from the query result, which does not meet the real-time constraints of an on-line search engine. It was used on IBM research model CLEVER [10].
We may assumed some graphs as input and all the above (PageRank, Weighted PageRank and HITS) algorithms are run on those graphs, then I have compared the results on the basis of output. We may implemented the formula to get the results graphically on MATLAB.

DOI : https://dx.doi.org/10.26808/rs.ed.i10v2.03
International Journal of Emerging Trends in Engineering and Development Issue 10, Vol.2 (Feb-Mar 2020)
Available online on http://www.rspublication.com/ijeted/ijeted_index.htm ISSN 2249-6149

PR(A)=(1-d)+d(PR(P1)/C(P1)+...PR(Pn)/C(Pn))

6. PROPOSED SOLUTION :PARTITIONING ALGORITHM

Because we know that page ranking algorithms work in an environment with hyperlinks. It is more suitable for parsing pages in the hyperlink area. Therefore, we first read the website in the proposed algorithm and then show it as a vector in the entire hyperlink area, however, since we know that due to its large number, it can form very large dimensional vectors of websites. To reduce the size and complexity, we only look at the first section of the URL (that is, from the URL http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/kmeans.html). We only take http: // home with .dei.polimi.it) during the first grouping loop and then re-group each cluster using the full URL. Now the pages are grouped into N groups, where N is the number of processing units. However, there is no guarantee that it converges to the global optimum. Therefore, it is repeated with different starting conditions until each group reaches an almost similar size (less than 10% standard deviation). The algorithm can be written in steps as follows: follow:

1. Represent each page by their hyperlinks
2. Now simultaneously count total no. of base urls from all Pages (suppose U)
3. Represent each page by U-dimensional vector
4. Perform k-means clustering for U groups, where U is the no. of total processing units(linked units)
5. Calculate the size of each cluster
6. If standard deviations of size of clusters are not less than 10%(threshold) then repeat from the step 4 with different initial values.
7. Now Count total no. of urls in one cluster (suppose U1)
8. Represent each page by U1-dimensional vector
9. Perform k-means clustering for N groups, where N is the no. of total processing units
10. Calculate the size of each cluster
11. If standard deviations of size of clusters are not less than 10% (threshold) then repeat from the step 9 with different initial values.
12. Calculate the distance among all cluster & place the clusters with minimum distance in same machine

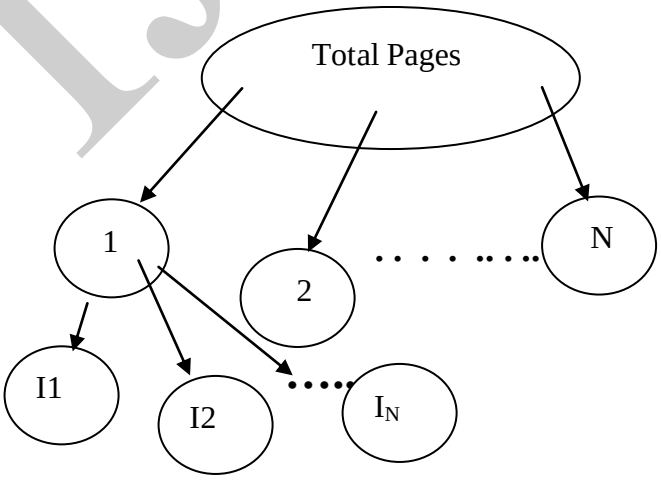

Figure 5 Showing Clustering approach

7. CONCLUSION

It has been analyzed in section 3 to 5, that a number of parameters are responsible for the web pattern analysis . But, it has been concluded that an algorithm with probable solution in polynomial time will become a solution of that problem. It is concluded here that this is NP-Complete problem and solvable in a polynomial Time. So, it needs to be solved using an algorithm, the algorithm thus proposed here named as partitioning algorithm which works on K-means. A thorough comparative analysis has been done in table 2 and section 4. In section 6, K- means based partitioning algorithm is used here to cluster the hyperlinking data and analyze the results. A large metahuristic data further clustered using it and the conclusion can be raised in clustered form. The objective of this paper was to give a solution to the problem and it has been given here in the form of an algorithm. This algorithm then will be analyzed in further research.

REFERENCES

[1] N. Grover, R. Wason, Comparative Analysis of Pagerank and HITS Algorithms,(IJERT),ISSN: 2278-0181,Vol.1 Issue8, 2012

[2] Olston, C. and Chi, E. H. ScentTrails: Integrating Browsing and Searchingon the Web. *ACM Transactions on Computer-Human Interaction (TOCHI),*Vol. 10, No. 3, pp. 177-197, 2003

[3] Perkowitz, M. and Etzioni, OAdaptive Web sites: automaticallysynthesizing Web pages. In *Proc. of AAAI'98*, pp. 727-732, July 26-30,Madison, Wisconsin, USA, ISBN 0-262-51098-7, AAAI Press, 1998

[4] Netcraft. Web server survey, 2004.

[5] Raymond Kosala, Hendrik Blockeel, Web Mining Research: A Survey,*SIGKDD Explorations, ACM SIGKD July* 2000.

[6] O.etzioni. The world wield web: Quagmire or Gold Mining.*Communicate of the ACM*, (39)11:65-68, 1996.

[7] S. Sharma, Saurabh, Dipti Jindal, and Rashi Agarwal. "An approach for congestion control in mobile ad hoc networks." *IJETED* , 2016.

[8] S. Sharma and R. Agarwal, "Optimizing QoS parameters using computational intelligence in MANETS," 2017 International Conference on Computing, Communication and Automation (ICCCA), Greater Noida, 2017, pp. 708-715, 2017

[9] Saurabh Sharma, Rashi Agarwal, Clustering and parameter optimization in MANETS using SOM and Genetic Algorithm, "" ARPN Journal of Engineering and Applied Sciences " Vol.13(23)pp. 9202-9212 , 2018.

[10] Wen-Chen Hu, Xuli Zong, Chung-wei Lee and Jyh-haw Yeh, World Wide Web Usage Mining Systems and Technologies.

[11]Nielsen, J. Designing Web Usability.*New Riders Publishing, Indianapolis,Indiana*, USA. , 2000

[12]Hong T, Chiang M, Wang S H, Mining weighted browsing patterns with linguistic minimum supports, *IEEE International Conference on Systems, Man and Cybernetics*, 2002,*Yasmine Hammamet, Tunisia*, pp. 635-639. 2002

[13] D. K Farkas,. and J. B Farkas,. Guidelines for Designing Web Navigation.*Technical Communication*, 47(3), pp. 341-358, August. 2000.

[14] L. Björneborn and P. Ingwersen, "Toward a Basic Framework for Webometrics". *Journal of the American Society for Information Science and Technology* 55 (14): 1216–1227, 2004.

[15] http://www.wisegeek.com

[16] Borgman, C & Furner, J. Scholarly communication and bibliometrics. In Cronin, B. (ed.), *Annual Review of Information Science and Technology 36* (pp. 3-72). Medford, NJ: Information Today Inc.. 2002

[17]Wouters, P. F. The citation culture. *Ph. D. thesis. University of Amsterdam.* (1999).

[18] Almind, T. C. & Ingwersen, P. Info metric analyses on methodological approaches to 'Webometrics.' *Journal of Documentation, 53*(4) 404-426. (1997).

[19] http://EzineArticles.com/

[20] http://www.eclickperformance.com/

[21] S. Brin and L. Page, The anatomy of a large-scale hypertextual web search engine,*Computer Networks and ISDN System*,30(1-7),,pp. 107-117,1998.