International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 3 | ISSN : 2456-3307

Page Rank Computation [PRC] in INformation Assimilation & Retrieval (INAR) System

Dr. L. Senthilvadivu
Principal, Mahendra Arts & Science College, Kalipatti, Tamil Nadu, India

ABSTRACT

The World Wide Web consists of millions of web pages and hyperlinks with which the most visited page and the number of hyperlinks in the page decide on the page rank. The page rank of a page is defined recursively and depends on the number and page rank metric of all pages that link to it. A page that is linked to by many pages with high page rank receives a high rank itself. In the INformation Assimilation and Retrieval (INAR) system, the page rank can be predictable based on the selection of the pages by the user who decide the page as the positive result of the search while inflowing the queries as search key from the heterogeneous and multi related information sources. The positive results of the users can be favored if and only if the page is the most relevant one for the queries posted by the user.

Keywords: - INAR, Assimilation, heterogeneous, queries

I. INTRODUCTION

Commonly the page rank of a web page can be calculated based on the importance of the particular page and its hyperlinks also. A Page Rank results from a mathematical algorithm based on the graph, the web graph, created by all World Wide Web pages as nodes and hyperlinks as edges. The rank value indicates an importance of a particular page. The Page Rank of a page is defined recursively and depends on the number and Page Rank metric of all pages that link to it. A page that is linked to by many pages with high Page Rank receives a high rank itself. If there are no links to a web page there is no support for that page. Page Rank optimization (PRO) is the process of improving the accomplishment of a website or a web page in search engines either organic or algorithmic search results. In general, higher ranked on the search results page, and more frequently a site appears in the search results list, the more visitors it will receive from the search engine's users.

II. INAR SYSTEM

For users of academic search engines, knowledge about applied ranking algorithms is also essential, for two basic reasons. Firstly, users should know about the algorithms in order to estimate the search engine's robustness towards manipulative attempts by authors and spammers and therefore, the trustworthiness of the results. Secondly, knowledge of ranking algorithms enables researchers to estimate the usefulness of results in respect to their search intention.

INAR system [2] consists of the powerful structure known as Assimilation and Search configuration development (**ASCD**) which contains the hierarchical persistence storage and combined data structure with Domain, Child Domain, Concept, Meaning, and Related Uniform Resource Locator (URL) with Serial number (S.NO) as Primary Key. Domain is the main concept; Child Domain is the related topic of Domain. Concept is most relevant

word of Domain and Child Domain. Meaning is the explanation and Related URL is the web page name that contains the comprehensive data. Information can be stored by the Administrator by himself and the information can be retrieved from the specified Servers. The default details of the INAR system are also stored by the Administrator. Administrator should authenticate with login page by selecting it from the Home. The home page has the relevant link to enter the contents which increases the heterogeneity.

Page rank can be calculated when the user visits the web page and the content of the page is pertinent hence the triumph of the user getting the more pertinent content is substantiated as the page rank by the user. The other user may find the other content as the page rank and this may be proceeding for different users, the rank can be calculated based on the frequency. The more number of users visit the page and select the same as the relevant page for the same query can be acquired for the computation of page rank of the meticulous page.

III. PAGE RANK COMPUTATION

Page rank can be computed with the support on the triumph over finding the most relevant page for the queries by the user and hence the users only authorize to decide whether the page is pertinent. The more no. of users visits the same page and select the meticulous page as the positive result of the search formulate the page rank of the page. Fig1 shows the structural design of page rank computation in INAR system. The users posted different queries in the form of single keyword, double keywords, and three keywords and so on to the system, which replies the content to the meticulous queries. The information is obtained from storage of the INAR system in addition to heterogeneous, multi related information sources label such as IS1, IS2 …ISn. The user while searching for the content by querying in the form of keyword perceptive that the correct page is found, the page can be entered as the positive result. The positive result is considered as the page rank of that page and it is intended with frequency distribution. When the same page is chosen by the other users for the same keyword, as the number of discrete events increases, the function begins to resemble a normal distribution.

PR = Max (v_1, v_2, v_3 ….v_n) by number of users (u_1, u_2, u_3… u_n)

Where PR is the Page Rank and v_1, v_2, v_3 ….v_n is the frequency of the page chosen by the number of users' u_1, u_2, u_3… u_n.

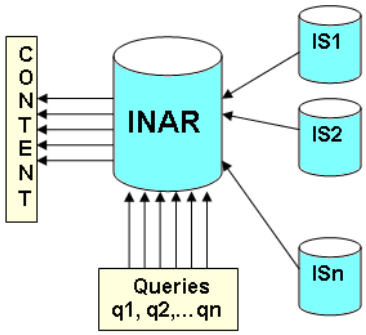

Fig 1 Structural design of Page Rank calculation

The calculation of the page rank in INAR system is done mathematically with the help of Gaussian function.

In mathematics, a Gaussian function is a function of the form:

$$f(x) = ae^{-\frac{(x-b)^2}{2c^2}}$$

For some real constants a, b, c > 0, and e ≈ 2.718281828 (Euler's number).The graph of a Gaussian is a characteristic symmetric "bell curve" shape that quickly falls off towards plus/minus infinity. The parameter a is the height of the curve's peak, b is the position of the centre of the peak, and c controls the width of the "bell".

The users may visit different pages furthermore number of times the identical page which can be

taken into account for the reckoning of the page rank. The computation of page rank of the selective page comprises of the number of times the page visited by the users and conclude the page is relevant for the meticulous query provided by the users.

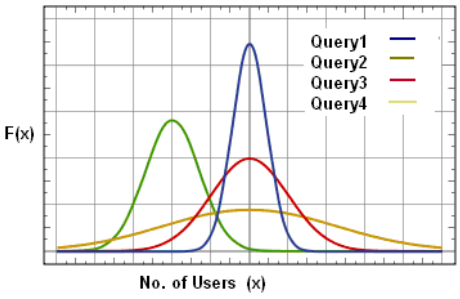

Fig 2 Page Rank Computation Curve

The curves are plotted between the no. of users and choosing the relevant pages for the queries F(x) and hence the maxima of the curves have decided the page rank of the substantial page or content.
Mathematically, the formula can be developed with the conditions that x = b and c = 1, the page rank can be

F(x) = a,

a is the height of the curve which is the maxima construe hypothetically. The curves are drawn for different queries posted by the users. Fig 2 shows the page rank computation curve which gives the maxima for different queries such as Query1, Query2, Query3 and Query4. Similarly, the curves are drawn for the different queries posted either by different users or the same users for different queries. Both the cases are incorporated for the computation of page rank of the relevant page. The Query1 is the word Physics, the corresponding result i.e., the relevant page for Physics is deemed as the page rank of the page for the Query Physics which is absolutely selected by the user who has given the query Physics. Similarly, the page rank can be computed by plotting the curve for other queries such as Query2, Query3, and so on. The maxima of these curve decides the page rank of the appropriate page which in addition concluding the relevancy of the page.

IV. RELATED WORK

Search Engine Optimization (SEO) may target different kinds of search, including image search, local search, video search, academic search,[1] news search and industry-specific vertical search engines.

In the present work, the INAR system gives the most relevant result for the computation of the page rank of the page which is decided by the user as the most germane for the individual queries posted by the user. The user is receiving the content directly in this system instead of searching with the millions of the links as in other systems for the particular query. More over the page is chosen as the most relevant by the maximum no. of users which computes the maxima of the frequency as the page rank.

As an Internet marketing strategy, SEO considers how search engines work, what people search for, the actual search terms typed into search engines and which search engines are preferred by their targeted audience. Optimizing a website may involve editing its content and HTML and associated coding to both increase its relevance to specific keywords and to remove barriers to the indexing activities of search engines. Promoting a site to increase the number of back links, or inbound links, is another SEO tactic.

A study about web search engines revealed that around 42% of all outgoing clicks were on the result [4]. Around 90% of outgoing clicks were on a result on the first page. That means that most users of a web search engine did not even pay attention to the second page. This illustrates the importance for webmasters to be listed in one of the very first positions and that ranking algorithms significantly influences the amount of visitors a webpage receives. It seems likely that the same is basically true for academic search engines.

The CLEVER search engine [14] incorporates several algorithms that make use of the Web's hyperlink

structure for discovering high-quality information. It can be exceedingly difficult to locate resources on the World Wide Web that are both high-quality and relevant to a user's informational needs. Traditional automated search methods for locating information on the Web are easily overwhelmed by low-quality and unrelated content. Second generation search engines have to have effective methods for focusing on the most authoritative documents. The rich structure implicit in hyperlinks among Web documents offers a simple, and effective, means to deal with many of these problems.

The criticism relates to the fact that citation measures impact but not quality in general [3, 4]. That means, articles with many citation counts are not always good. It might make sense to rank articles with high citation counts first if a user is searching for standard literature with high impact. But there may be situations in which it makes no sense to display highly cited papers in the first positions. This could be, for instance, if someone searches for the latest trends in a certain research field or articles from authors advancing a view different from the majority.

Several studies about Google Scholar exist. Studies include, for instance, research into data overlap with other academic search engines such as Scopus and Web of Science [5], [6], Google Scholar's coverage of the literature in general and in certain research fields [7], [8], the suitability to use Google Scholar's citation counts for calculating bibliometric indices such as the h-index [9] and the reliability of Google Scholar as a serious information source in general [10], [11]. Google Scholar itself publishes only vague information about its ranking algorithm: Google Scholar sorts "articles the way researchers do, weighing the full text of each article, the author, the publication in which the article appears, and how often the piece has been cited in other scholarly literature" [12]. Any other details or further explanation is not available.

Although Google Scholar's ranking algorithm has a significant influence on which academic articles are read by the scientific community, we could not find any studies about Google Scholar's ranking algorithm despite our own one [13].

A web surfer who chooses a random link on every page (but with 15% likelihood jumps to a random page on the whole web) is going to be on Page E for 8.1% of the time. (The 15% likelihood of jumping to an arbitrary page corresponds to a damping factor of 85 %.) Without damping, all web surfers would eventually end up on Pages A, B, or C, and all other pages would have Page Rank zero. Page A is assumed to link to all pages in the web, because it has no outgoing links.

V. CONCLUSION AND FUTURE WORK

The INAR system is the appropriate system to compute the page rank with more accuracy. The information retrieval in this system is not based on the links in which again the users have to mine for their information as in other systems. This makes the users enervate to burrow for the web again and again but in vain. The INAR system provides the relevant content for the query within a smaller amount of time. The page rank can also be computed based on the relevancy of the page, i.e., the relevancy is experienced by more number of users depending on the number of times the page is chosen for the same query, the page rank is preferred. In future, this system can be enhanced with adding the dictionary to the system for discriminating the countless meanings for the same utterance.

VI. REFERENCES

[1]. Beel, Joran and Gipp, Bela and Wilde, Erik (2010). "Academic Search Engine Optimization (ASEO): Optimizing Scholarly Literature for Google Scholar and Co.". Journal of Scholarly Publishing. pp. 176–190.

[2]. L. Senthilvadivu, K. Duraiswamy (2011) "Conniving the Information Assimilation and Retrieval (INAR) system for the heterogeneous, multi related Information Sources", World of Computer Science and Information Technology Journal pp.357-363.

[3]. Jon Kleinberg (1999) "Mining the link structure of the World Wide Web and Trawling the Web for emerging cyber-communities" Eighth World Wide Web conference, Toronto, Canada

[4]. (2006, August) Click through Rate of Google Search Results - AOL-data.tgz - Want to Know How Many Clicks The no.1 Google Position Gets? Red Cardinal Blog. Red Cardinal Ltd.

[5]. J. Bailey, C. Zhang, D. Budgen, M. Turner, and S. Charters (2007), "Search Engine Overlaps: Do they agree or disagree?" in Second International Workshop on Realizing Evidence-Based Software Engineering (REBSE '07), p2.

[6]. K. Yang and L. I. Meho (2006) "Citation Analysis: A Comparison of Google Scholar, Scopus, and Web of Science," in 69th Annual Meeting of the American Society for Information Science and Technology, Austin (US), pp. 3–8.

[7]. W. H. Walters (July 2007), "Google Scholar coverage of a multidisciplinary field," Information Processing & Management, vol. 43, no. 4, pp. 1121–1132.

[8]. J. J. Meier and T. W. Conkling (2008), "Google Scholar's Coverage of the Engineering Literature: An Empirical Study," The Journal of Academic Librarianship, vol. 34, no. 34, pp. 196–201.

[9]. J. Bar-Ilan (2007), "Which h-index? - A comparison of WoS, Scopus and Google Scholar," Scientometrics, vol. 74, no. 2, pp. 257–271.

[10]. P. Jacso (2005), "Google Scholar: the pros and the cons," Online Information Review, vol. 29, no. 2, pp. 208–214.

[11]. B. White (2006), "Examining the claims of Google Scholar as a serious information source," New Zealand Library & Information Management Journal, vol. 50, no. 1, pp. 11–24.

[12]. (2008) About Google Scholar. Website. GoogleInc.OnlineAvailable:http://scholar.google.com/intl/en/scholar/ about.html

[13]. J. Beel and B. Gipp, (2009) "Google Scholar's Ranking Algorithm: An Introductory Overview.

[14]. Jon Kleinberg (2011), IBM's Almaden Research Center