Interaction Information Retrieval and Improved Page Rank Algorithm Based on Access Duration of Page

Shail K Dinkar
Assistant Professor, Department of Master of Computer Application, G. B. Pant Engg. College,Pauri Garhwal-246194,India

Hemant Kumar
Assistant Professor, Department of Mechanical Engg.,G. B. Pant Engg. College, Pauri Garhwal-246194,India

Abstract

The PageRank method is used by the Google Web search engine to compute the importance of Web pages. Two different views have been developed for the interpretation of the PageRank method and values: (a) stochastic (random surfer): the Page Rank values can be conceived as the steady-state distribution of a Markov chain, and (b) algebraic: the Page Rank values form the eigenvector corresponding to eigenvalue 1 of the Web link matrix. The Interaction Information Retrieval (I^2R) method is a non-classical information retrieval paradigm, which represents a connectionist approach based on dynamic systems. In the present paper, a different interpretation of PageRank is proposed, namely, a dynamic systems viewpoint, by showing that the PageRank method can be formally interpreted as a particular case of the Interaction information Retrieval method;

Keywords: Page Rank,Web Graph , Age of Page, Node, Damping factor

1.Introduction

The PageRank method (Brin. & Page. 1998) [1] is an important component of the Google Web Information Retrieval (IR) engine (Google), and triggered research into and application of link analysis to the Web. It allows the calculation of a priori importance measures for Web pages. The measures are computed offline, and are independent of the search query. The PageRank values form the eigenvector corresponding to eigenvalue 1 of the Web matrix. At query time, the measures are combined with query-specific scores used for ranking Web pages.

The PageRank method uses the metaphor of an easily bored Web surfer. The PageRank value of a page is conceived as being the probability that the surfer reaches that page by following forward links. The PageRank values form a probability distribution over the Web. The PageRank algorithm has been intuitively justified to model a random surfer in which a user clicks on the links at random and the rank of a page signifies the probability of the user arriving at that page. This algorithm does not consider the time constraint (age of the specific page on the web). It is observed that a number of incoming links which point at a page (with same content) may increase with time because when different

persons become aware of the importance of the web page with time, they provide a link to that specific page, thus the number of incoming links to that particular web page may increase with time. Therefore, time must be considered to decide the priority (rank) of the page. Hence time constraint is introduced in the existing algorithm to derive a new and effective algorithm.

2. PageRank

The Google search engine exploits the citation graph of the crawled portion of the publicly accessible World Wide Web (briefly Web), and calculates a measure of the relative importance of Web pages using a stochastic process view

Place of PageRank in Google's Retrieval and Ranking

The retrieval and ranking of Web pages follows a usual IR scenario, and is performed in several steps as follows:

1 Find the Web pages containing the query terms.
2 Compute a relative importance of Web pages.
3 Rank the Web pages according to their relative importance.

The relative importance of Web pages is calculated taking into account several factors such as:

- "On page factors," i.e., terms occurring in title, anchor, body, proximity of terms
- Appearance of ,temis: small font, large font, color
- Frequency of occurrence of terms

- PageRank values
- Other factors

Although not known publicly exactly, based on (Brin & Page. 1998) [1] it can be assumed that two numeric vectors (for query and Web page) are delined, whose dot product gives an intermediate score for that Web page, which is then combined with the PageRank value of that Web page to obtain its final score (importance).

2.1 The Page Rank Method

The PageRank method is considered to be one of the factors used by Google in computing the relative importance of Web pages. The PageRank value of a Web page depends on the PageRank values of pages pointing to it and on the number of links going out of these pages.

The principle: The starting point for the principle of PageRank is citation analysis, which is concerned with the study of citation in the scientific literature, and dates back, in its present form, about 50 years (Garfield. 1955) [2]. The underlying idea of citation analysis—which is well-known and highly published—reads as follows: citation counts are a measure of importance (Garfield. 1972) [3]. This idea was used by Caniere and Kazman (1997) [5] for Web retrieval. In the PageRank method (Page, Brin. Motwani, & Winograd. 1998)[4], this idea is refined in that citation counts are not absolute values anymore, rather relative ones and mutually dependent (as will be shown below). The principle on which PageRank is based can thus be referred to as an extended citation principle, and can be formulated as follows: a Web page's importance is determined by the importance of Web pages linking to it.

The model: To apply the extended citation principle in practice, an appropriate model of a system of Web pages has been constructed as follows. Let (Figure 1)

(i) $\Omega = \{W_1, W_2, \ldots, W_i, \ldots, W_N\}$ denote a set of Web pages under focus,
ii) $\Phi_i = \{W_k | k = 1,\ldots,n_i\}$ denote the set of Web pages W_i, points to, $\Phi \subseteq \Omega$
(iii) $B_i = \{W_j | j = 1,\ldots,m_i\}$ denote the set of Web pages that point to W_i, $B_i \subseteq \Omega$
It can be seen that this view models the Web as a directed graph denoted by, say, *G.*

i .
B Φ_i
W_j W_i W_K

FIG. 1.Web pages and links are viewed as a graph G in PageRank, where pages W_1, W_2, ….., W_i, ….., W_N are its nodes.
$B_i = \{W_j | j = 1,\ldots,m_i\}$ denote the set of Web pages that point to W_i, $\Phi_i = \{W_k | k = 1,\ldots,n_i\}$ denote the set of Web pages W_i, points to.

The equation: Based on the extended citation principle and using the graph model. Haveliwala (1999) [6] and Page (2001) [7] define the PageRank value of a Web page W, denoted by *Ri*, using the following equation:

$$P(R_i) = \sum_{W_j \in B} \frac{P(R_j)}{L_j} \quad (1)$$

where L_j denotes the number of outgoing links from the page W_j. Equation 1 is a homogenous and simultaneous system of linear equations, which is always known to have trivial solutions (the null vector), but which has nontrivial solutions too if and only if its determinant is equal to zero.
Another commonly used technical definition of Page-Rank is as follows. Let G = (V, *A)* denote the directed graph of the Web. where the set $V = \{ W_1, W_2, \ldots\ldots, W_i, \ldots ., W_N\}$ of vertices denotes the set of Web pages. The set *A* of arcs denotes the links (given by URLs) between pages. Let $M = (m_{ij})_{NXN}$ denote a square matrix attached to graph G such that $m_{ij} = 1/L_j$ if there is a link from W_i to W_j. and 0 otherwise. Because the elements of the matrix *M* are the coefficients of the right hand side of Equation 1, this can be rewritten in a matrix form *as M X R = R*, where *R* denotes the vector of Page Rank values, i.e., $R = [R_1, \ldots\ldots, R_i, \ldots ., R_N]^T$

If the graph G is strongly connected, the column sums of the matrix *M* are equal to 1 (stochastic matrix). Because the matrix *M* has only zeroes in the main diagonal, the matrix *M -1* has zero column sums (I denotes the unity matrix); i.e., the matrix obtained by subtracting 1 from the main diagonal of the matrix M). Let *D* denote its determinant, i.e., $D = |M-1|$. *If* every element of, say, the first line of *D* is doubled we get a new determinant *D',* and we have *D' = 2D.* We add now, in *D',* every other line to the first line. Because the column sums in *D* are null, it follows that *D' = 2D = D,* from which we have *D = 0.* The matrix *M — 1 is* exactly the matrix of Equation I. hence it has nontrivial solutions too (of which there are an infinity). The determinant $|M-1|$ being equal to 0 is equivalent to saying that the number 1 is an eigenvalue of the matrix *M.*

The PageRank values are computed in practice using some numeric approximation procedure to calculate the

eigenvector *R* corresponding to the eigenvalue 1 or to solve Equation 1. Computational details as well as convergence considerations in the numeric algorithms used are presented and discussed in Arasu (2002) [10]. Haveliwala (1999. 2002) [6,8]. Kim and Lee (2002) [9]. Figure 2 shows an example.

As it was initially suggested (Brin & Page. 1998), the PageRank value of a Web page can be interpreted using an easily bored random surfer metaphor: the surfer clicks on links at random with no regard towards content. The random surfer visits a Web page with a certain probability, which derives from the page's PageRank. The probability that the random surfer clicks on one link is solely given by tbe number of links on that page. This is why one page's PageRank is not completely passed on to a page it links to, but is divided by the number of links on the page. The probability for the random surfer reaching one page is the sum of probabilities for the random surfer following links to this page; his probability is reduced by a damping factor *d*, set between 0 and 1. The justification is that the surfer does not click on an infinite number of links, but gets bored sometimes and jumps to another page at random. The higher *d* is, the more likely will the random surfer keep clicking links. Since the surfer jumps to another page at random after he stopped clicking links, the probability therefore is usually implemented as a constant *(1-d)* into the numeric approximation algorithm. A PageRank of zero means the page has no reputation at all. All PageRanks sum up to 1, since after *n* clicks, the surfer must he on exactly one web page. Thus, all the PageRank values may be interpreted as forming a probability distribution over the Web, so that the probabilities sum up to unity. The solutions of Equation 1 can be so scaled as to satisfy this condition.

2.2 Description of PageRank Calculation

Academic citation literature has been applied to the web, largely by counting citations or back links to a given page. This gives some approximation of a page's importance or quality. PageRank extends this idea by not counting links from all pages equally and by normalizing the number of links on a page.

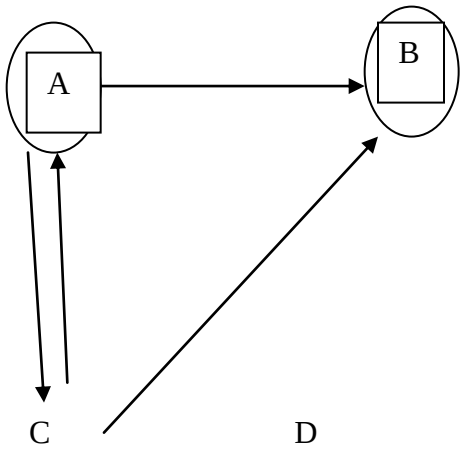
The PageRank algorithm was intuitively justified to model a random surfer in which a user clicks on links at random and the rank of a page signifies the probability of a user arriving at that page. A user can arrive at a page either by clicking on the links or by randomly jumping to a page. The algorithm includes a parameter d, which represents the probability of a user continuing to click on links and (1–d) as the probability that the user jumps to a random page. PageRank of a page is determined using the random surfer model described
above. The PageRank of a page A can be computed [11,12] as follows:

$$PR(A) = \frac{1-d}{N} + d * \sum for\ every\ j \in s \left(\frac{PR(j)}{L(j)} \right)$$

where,
PR(j): is the PageRank of page j.
S : is the set of nodes that have a link to page A.
L(j) : is the out degree of page j.
d : is the damping factor that is set to a value between 0 and 1. It is usually set to 0.85 for the web graph.
N : is the number of nodes in the graph.

Addition of d also addresses the issue of a node being a rank source, i.e., a node having zero in links will be assigned at least a minimum PageRank value of d. In order to overcome the problem of a rank sink, i.e., a node having zero out links, it is assumed that there exists an iterative computation. The PageRank of all the nodes in the graph is computed using an iterative algorithm. Each node is assigned an initial value and the PageRank of all the nodes is then calculated in several iterations using an equation-based on PageRank algorithm [11] shown in Figure 2.

In the PageRank algorithm, a page uniformly distributes its rank to each outlink and in turn



Iterative Computation would be

PR(A) = (1 – 0.85)/4 + 0.85 * (PR(C))
PR(B) = (1 – 0.85)/4 + 0.85 * (PR(A)/2)
PR(D) = (1 – 0.85)/4
PR(C) = (1 – 0.85)/4 + 0.85 * (PR(A)/2) + PR(B) + PR(D)

Figure 2: Example of Computation of PageRank Algorithm

when computing the PageRank of a page,the rank of each inlink is weighed and equally outlinked to all other nodes in the graph, and there is an increased probability of starting at a random page.

3. Drawback of the Existing Algorithm

In the existing algorithm, time factor is not considered for calculation of PageRank.We know that if any web page is rich in C D terms of content, that specific page may be pointed by large number of different pages. Iterative Computation would be If any web page is available on the web for a long time, then the number of incoming links increase with time. So for a given page, the number of links pointing to a specific page may increase with time. If pages A and B are supposed to have the same contentwise weight, and if page A is older than page B, then more pages may point to page A of PageRank Algorithm compared to page B on the web. Therefore, time factor must be considered for defining the PageRank of any web page.

4. Methodology

New Improved Ranking Algorithm of Web Pages in Search Engine The existing algorithm does not consider the time constraint (age of that specific page). It is observed that a number of incoming links pointing to a particular page (with the same content) will increase with time, so time must be considered to decide the priority of the page. New optimized algorithm is obtained by introducing the concept of time constraint (age of that specific page). If the current ranking algorithm is modified and rank per unit time is found, then it is justified.
In the current scenario of search engines, when we search for a specific topic, the relevant pages are retrieved on the basis of priority. By taking a directed graph where a node represents the web page, we modify the ranking algorithm by introducing time constraint as follows:

Procedure: Improved_ALGO (in G: Directed Graph with N nodes.N: N is the number of nodes in the graph,)

PR[1..N]: Rank values of the nodes per unit time

t[1..N]: Age of nodes

d: Dampening factor. It is usually set to 0.85 for the web graph

C(j): Out degree of page j.

S: Set of nodes that have an in link to page i

t(i): Age of page i

for (i = 1 to N)

find t(i)

[End of for loop]

for (i = 1 to N)

// calculates PageRank of page i per unit time // for finding PageRank per unit time, PageRank

PR(i)={(1–d)/N+d * for e v ery j ((PR(j)/C(j))}/t(i)

// S: set of nodes that have an in link to page i

// PR(j) :is the Page Rank of page j

// C(j) : out degree of page j.

[End of for loop]
Exit

5. Results and Discussion

The key findings and their implications from the above improved search algorithm is as follows:
We are also considering the time constraint (age of the specific web page on the web) in the modified improved algorithm to define the PageRank for appropriate results. If a valuable web page is

recently registered to the global web (age of web page on global web is low), then the number of incoming links pointing to it will also be low. Those web pages whose age is high may have larger number of incoming links due to more availability on the web although importance of web page is identical to lower age page with respect to other aspects like content. But if we consider the time (age) of web page with respect to its availability on the web for ranking, then the average incoming link per unit time can give efficient results because page rank is calculated on the basis of per unit time in improved page ranking algorithm.

Conclusion

The Web has become 'the place' for accessing any type of information. There are billions of web pages and everyday new content is produced. Therefore, the use of search engines is becoming a primary Internet activity, and search engines have developed increasingly clever ranking algorithms in order to constantly improve their quality. Nevertheless, there are still many open research areas of tremendous interest where the quality of search results can be improved. The paper proposes a highly effective and efficient solution for search result records from search engine response pages, by considering the time constraint (age of the page) and the number of incoming links to a particular page. Thus, we can say that by applying the above modified algorithm, search quality can be improved.

References

1. Brin S and Page L (1998), The Anatomy of a Large-Scale Hypertextual Web Search Engine_, Computer Networks and ISDN Systems, Vol. 30, Nos. 1-7, pp. 107-117.
2. Garfield, E. (1955). Citation indexes for science.Science 122, 108- 111.
3. Garfield, E. (1972). Citation indexes for science. Science, 471 - 479.
4. Brin, S., Motwani, R., Page, L., & Wingrad, T. (1998). What can you do with a web in your pocket? Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, 21(2), 37 – 47. Retrieved November 4, 2002, from http://www.n3labs.com/pdf/brin98what.pdf.
5. Carriers, J., & Kazman, R. (1997, April), Webquery: Searching and visualizing the web through connectivity. Computer Networks, 30 (1-7), 107-117.
6. Haveliwala, T.H (1999). Efficient computation of pagerank (Stanford University). Retrieved November 4, 2002, fromhttp://dbpubs.stanford.edu:8090/pub/1998-31.
7. Page, L. (200). United States Patent 6.285.999. September 4. Retrieved November 4, 2002, from http;//164.195.100.11/netacgi/…/6.285.999.
8. Haveliwala, T. H. (2002, May). Topic-sensitive pagerank. Proceedings of the World wide Web Conference 2002, Honolulu, HI. Retrieved November 4, 2002, from http://www2002.org/CDROM/
9. Kim, S.J., & Lee, S.H. (2002). An improved computation of the pagerank algorithm. In F. Crestani, M. Girolamo, & C.J. van Rijsbergen (Eds.), Proceedings of the European Colloquium on Information Retrieval (LNCS 2291. Pp. 73-85). London: Springer.
10. Arasu, A. (2002, May). PageRank computation and the structure of the web: Experiments and algorithms. Proceedings of the World Wide Web 2002 Conference, Honolulu. HL. Retrieved November 4, 2002, from htto://www2002.org/CDROM/poster
11. Padmanabhan D, Desikan P, Srivastava J et al. (2005), Web Intelligence, 2005 Proceedings_, The 2005 IEEE/WIC/ACM International Conference on Web Intelligence, Copiane University, France, September 19-22.
12. Padmanabhan D, Desikan P, Srivastava J et al. (2005), ICER: A Weighted Inter-Cluster Edge Ranking for Clustered Graphs_, Web Intelligence 2005 Proceedings, The 2005 IEEE/WIC/ACM International Conference on Web Intelligence, Copiane University, France, September 19-22.