ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, *Vol 9(48), DOI: 10.17485/ijst/2016/v9i48/107962, December 2016* ISSN (Online) : 0974-5645

Proximity Prestige using Incremental Iteration in Page Rank Algorithm

M. Anuradha* and J. Sethuraman

School of Computing, SASTRA University, Thanjavur - 613401, Tamil Nadu, India;
anuvendra@gmail.com, sethuraman@cse.sastra.edu

Abstract

Background/Objectives: Search engines such as Google uses the indexed results to respond to the user's query. It also uses ranking algorithm namely Pagerank algorithm to rank the web pages. Ranking mechanism works offline and the value is static. Pagerank algorithm uses uniform probability distribution and Power iteration to rank the web. Though the method is simple and efficient, applying it to large scale is not effective since the computational cost is expensive and the convergence occurs at slower rate. **Methods:** The proposed work Proximity Prestige algorithm uses degree prestige along with proximity prestige and the calculation method used is Incremental Iteration. In Degree Prestige, a web page has more prestige if it has many in links. Proximity Prestige is defined by the closeness of other web pages that links to that page. It uses non-uniform transition probability for computing the rank vector. **Findings:** This work enhances the rank of a web page. From this work it is proven that when a page is more close to its home page and if it receives more in links, its rank value increases. **Applications:** It can be used by the search engines to compute the rank of a web page in accordance with the proximity and the incremental iteration method could be applied in the area where computational cost is expensive.

Keywords: In Links, Matrix, PageRank, Prestige, Transition Probability

1. Introduction

In the growing trend of the web, text-based search is not sufficient to meet the query needs of the user. This gave rise to a static method computed offline which ranks the web pages based on the quality of in links, namely PageRank. It specifies the probability of each user at that node. A higher PageRank gets more importance and that web page gets listed when a search for that relevant topic is done. Search engine uses ranking algorithms such as Pagerank and Hits algorithm to rank the web pages. It is one of the criteria in selecting the most relevant web page for a user query. PageRank algorithms are widely used in social networks, object databases and recommendation systems, citation works, author rankings and in distributed networks. Google's PageRank is based on the fact that hyperlink represents a vote of one page for the other. From the hyperlink structure, the algorithm decides the popularity of a page. Most methods use this link structure and eigenvector to handle the hyperlink structure and rank score. PageRank algorithm performs uniform distribution of the rank scores to the pages to which it links, even though they are not important. The rank vector is computed iteratively by using Power iteration method which computes the eigenvector of the given matrix, i.e. hyperlink matrix. Web mining is a domain which helps to retrieve useful information from the web. It is of three types such as web content mining, web structure mining and web usage mining. Web usage mining is a process that extracts information from logs of server and browser cookies. Web content mining is the text and multimedia data found in a web page. Web Structure mining is basically a web graph which structures nodes as web pages and links as hyperlinks. The structural representation is found in the Figure 1. Hyperlinks are of two types inlinks and outlink. Inlinks or back links are links that are received by a web page from another web page. Outlinks are links that point to another page from the current one. The more backlinks a webpage has, it accumulates more rank and more authority. The entire web is visualized as a graph, where the nodes denote pages and edges represent links. The graph is

* *Author for correspondence*

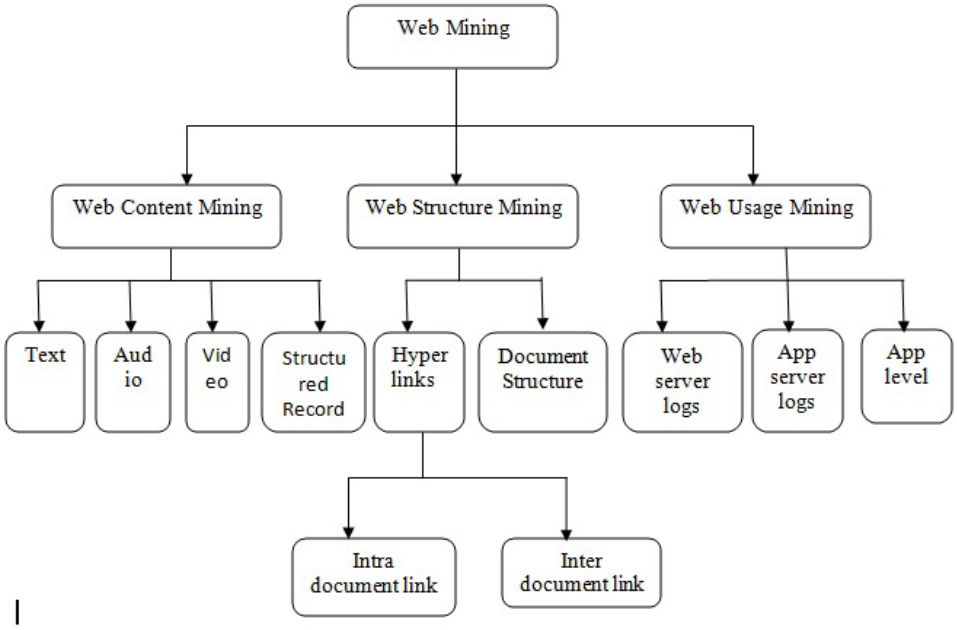

Figure 1. Web mining classification.

symbolized by a hyperlink matrix. It represents the whole web structure and the web size is huge in general, so a fast calculation method is required to efficiently compute a large number of page ranks on the web. Several algorithms are used for rank computations. Power iteration is the basic and conventional method. Other methods include Incremental iteration method, Distributed page rank. PageRank calculation methods use directed-link matrix and its eigenvector to handle hyperlink structure and ranking score. With a given hyperlink matrix, PageRank calculates its vector by using an iterative multiplication procedure, to compute dominant eigenvector of a given matrix. And if a page receives a rank from most valuable page such as wiki or Google, then it is assumed to get more weightage. PageRank algorithm alone doesn't determine the importance of a web page. The proposed algorithm works by taking into account the degree prestige and proximity prestige of a web page. And the method works by using incremental iteration as computation method. It uses a novel method namely Incremental Iteration. In this method if already calculated components of PageRank vector are immediately used to compute uncalculated components, its convergence speed can be improved. A key idea of our method is to immediately use previously calculated components when calculating the next (uncalculated) components of PageRank vector. And also Proximity Prestige Rank value is considered while calculating the rank value. Proximity prestige, a web page is defined by the distance or closeness of other pages that link to it.

1.1 Rank Sink

In a web graph, multiple pages may connect with each other. They build a loop themselves and they do not distribute rank to other pages. Due to this, pages in the loop receive more Page Rank values than existing. In Figure 2, nodes 1, 2, 3 distribute their ranks and node 5 receives the rank value but does not share its rank. So rank sink occurs at this node.

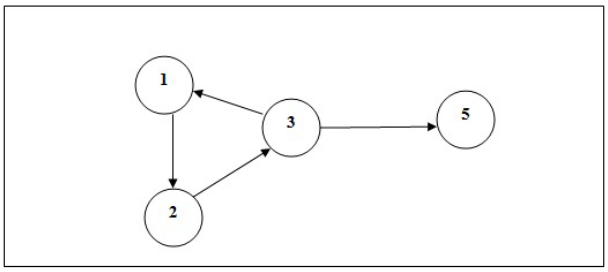

Figure 2. Rank sink representation.

1.2 Damping Factor (d)

It is the probability a random surfer follows a link. The random surfer gets bored and doesn't follow the link and has the probability of 1 – d. The damping value 'd' exists

in the range 0 and 1, analysis is done for the values 0.5, 0.85. Search engine make use of Pagerank algorithm to rank the web pages. By optimizing this algorithm user gets more relevant results for his query. The proposed work uses degree prestige along with proximity prestige and the calculation method used is Incremental Iteration. Degree prestige of a page is defined as the number of inlinks it receives and Proximity prestige defines a web page as prestigious if it is closer to its home page. This work enhances the rank of a web page. When a page has higher rank then it is said to have more probability of being selected by the search engine and gets selected by the user and the quality of ranking could be improved.

This paper is described as follows: In Existing Work section, rank calculation methods are discussed. In Discussion and Evaluation section the problem statement along with the proposed algorithm and the experimental values are detailed. In Conclusion section results obtained and the future scope is discussed. In References section all the cited works are quoted.

1.3 Existing Work

In[1] proposed Pagerank algorithm, which works on search and analysis of link in web. Search engine is designed using this algorithm and it is independent of query. Initial Probability Distribution (IPD) and Transition Probability Distribution Matrices (TPD) are the core metrics for determining the rank value. Most of the work is in designing the values of these matrices. In[1] states that every page which has same probability has to be taken as initial point. In[2, 3] suggests that the initial probability distribution vector is 1/N, where N refers to the number of pages. Since hyperlinks are selected uniformly transition probability matrix is created and the value is 1/outbound links. A page k has outlinks to N web pages and page l is one of it, then the probability of visiting page l from page k is 1/N[4]. A page without an outgoing link is said to be a dangling page[5]. In the existing work, dangling nodes are excluded from the calculations and rank value is computed. Later dangling pages are added to the system. TPD matrix is a stochastic matrix in which each column of the transition probability matrix sums to one and it is a non negative real numbers. If the transition probability matrix has dangling column, then it is said to be not a stochastic matrix. In all the above papers 1/N is taken as IPD value and value for the transition probability. In[6], a web page is said to be strongly connected, when there exists an edge from every page to other page. A random surfer may click uniformly by random to reach next page. PageRank is said to work for directed graphs. Most of the Pagerank methods use link matrices of a directed graph and its eigen vector to represent the hyperlink structure of the web and to calculate the rank. With the matrix, Power iteration method could be used to iteratively is calculate the eigen vector of a matrix. Since the size of matrix is high, applying power iteration could be time consuming and inefficient. Hence Incremental iteration method could be used to calculate the rank vector. In[7] proposed the incremental iteration that reduces the computational cost and convergence could occur at a faster rate. By applying the method pagerank calculation is done much faster by reducing the computational cost. Experiments are conducted by applying the power iteration and with incremental iteration and the results are verified with correlation coefficient to check the accuracy. It[8] addresses the problem of dangling page by introducing a solution namely teleport operation. In this a random surfer navigates from dangling page to other pages. Transition probability is said to be 1/N. In[9] suggested a proportionate prestige method to calculate the proportionate transition probability matrix with degree prestige and power iteration method. It overcomes the dangling issue. It boosted up the most relevant web page and drags down the least significant page. In[10] Search engines are optimized using fuzzy techniques. The parameters include pagerank, mouse and eye movements of user by assuming search engine as a user to determine the search results. Ontology method[11] is created for user profile for ranking. Based on user's click user's preferences are accessed. This helped the search engines to increase the performance by using profiles of the user.

2. Proposed Architecture

As in Figure 3, user searches for a query through search engine. Crawler is a software that spiders the entire web and keeps the index of results based on certain keywords. When search engines receive an input from the user it gets the information retrieval score from the crawler. It gets the corresponding rank of the web pages from the Pagerank algorithm. The algorithm receives web page as input and their corresponding degree and proximity prestige and calculates the rank value.

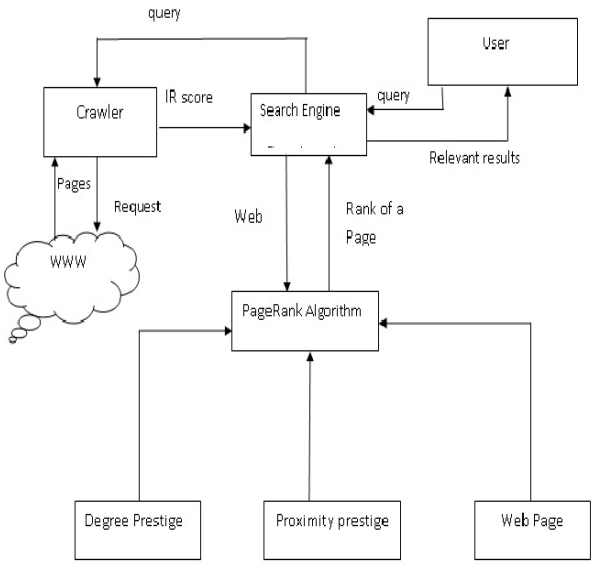

Figure 3. Pagerank computation – process.

3. Discussion and Evaluation

3.1 Rank Calculation - Uniform Probability Distribution (UPD) using PageRank algorithm

Consider web as a directed graph G = (V, E). V denotes vertices and E represents edges. Adjacency matrix denoted as $A_{ij.}$

$A_{jk} = \{ 1$ if $(j, k) \in E$ (1)
0 else
}

Page Rank of page k is given by,

$PR(k) = \Sigma_{(j,k)\in E} PR(j) / Out_j$ (2)

Transition Matrix TPD matrix, C is given by

$C_{jk} = \{1/ Out_j$, if $(j, k) \in E$ (3)
0 else
}

Every entry of transition probability matrix lies in the interval 0 to 1 and columns and should be of norms 1. Initial Probability Distribution (IPD) values is a vector. It is given by pr_0 (1), pr_0 (2), ……. pr_0 (N) ^{T} which is a column vector and it should satisfy Equation (1)

The probability of a surfer that to be in page j after one transition is obtained by using

PR_1 (j) = x PR_0 (i) (4)

As a result, the transition probability is given by,

$PR_{n =} C$ x PR_{n-1} (5)

The equation represents the characteristic equation of eigen system where the solution to Pagerank is an eigenvector with the eigen value equal to 1. So power iteration is used to find the page rank.

3.2 Rank Calculation - Non Uniform Probability Distribution (NUPD) using Proximity Prestige Algorithm

Initial probability matrix of each page is calculated using the inlinks of each web page. Each value is computed using the formula a/b where a represents inlink of each page and b represents the total number of links in each graph. It satisfies the Equation (6),

$\Sigma_{i=1 \text{ to } N} PR_0(i) = 1$ (6)

And the value lies between 0 and 1.

Prestige (P_i)of every page is computed by summing up the weight of inlinks of each page.

$P_i = \Sigma_{j=1 \text{ to } N} C_{ij}$ (7)

If a page has multipled links and Total Prestige is obtained by adding the prestige of those pages.

$TP_j = \Sigma_{j=1 \text{ to } N} P_i$ (8)

Each value of transition probability matrix is derived using the following formula - Each page's prestige divided by the Total Prestige.

$P_{ij} = P_i / TP_j$

Proximity of each page is the parameter value that defines the closeness from its home page.

3.3 Proximity Prestige Algorithm

Input: Set of pages with urls, connectivity of each web page

Output: Rank of a page

Process:

- Load the dataset into the program and process it to find the proximity prestige.
- Find the adjacency matrix with the dataset and find the initial probability matrix for each page such as a/b.
- Formulate the transition probability matrix C using the formula P_i/TP_j.
- Calculate Prestige of each page P_i using uniform probability distribution.

- Find $TP_{j,}$ total prestige if a web has more than an outbound link.
- Compute $P_{ij,}$ the prestige score which accommodates degree prestige and proportionally distributed.
- For each page add the proximity prestige value that is calculated separately.
- If all the values of a page are zero, then replace it with the proportional value of that page.
- Incorporate the damping factor by multiplying it with 1-d, where d is damping factor and the value of d exists between 0 and 1.
- Add d/N value to each entry in the matrix, where N refers to the number of pages. Resultant matrix is the transition probability matrix C.
- With the IPD and TPD values, use incremental iteration to derive the rank scores.

3.4 Incremental Iteration

In incremental iteration, *i*'th component of PageRank vector at *k*+1' th iteration is calculated as follows

$$r_i^{k+1} = {}_j^{k+1}m_{ji} + {}_j^{k}m_{ji} \quad (9)$$

In this method if already calculated components of PageRank vector are immediately used to compute uncalculated components, its convergence speed can be improved. A key idea of this method is to immediately use previously calculated components when calculating the next (uncalculated) components of PageRank vector. The values that are obtained by using existing Uniform distribution and Proposed Proximity Prestige for a sample of 20 pages are detailed in Table 1. It shows for most of the pages the rank value increases and the comparison is represented by the graph. Red bar in the Figure 4 denotes the existing UPD values that are obtained using the existing PageRank algorithm and Blue bar in the chart denotes the values that are obtained using Proximity prestige values. The graph clearly shows for the pages which has degree prestige value higher and which are in proximity to their home pages has values higher than the other web pages.

Table 1. Comparison of rank values

Pages	UPD values	Proposed Proximity Prestige values
1	0.045600	0.0376915
2	0.045925	0.0480545
3	0.035858	0.0388475
4	0.036549	0.0403165
5	0.105516	0.0907469
6	0.055802	0.0807300
7	0.031406	0.0280545
8	0.041144	0.0519960
9	0.051379	0.0535835
10	0.071830	0.0735835
11	0.038951	0.0449425
12	0.048693	0.0403505
13	0.066335	0.067139
14	0.029569	0.0288150
15	0.048573	0.0462945
16	0.051251	0.0531575
17	0.035858	0.0424765
18	0.042130	0.0397230
19	0.074198	0.0720350
20	0.043433	0.0462880

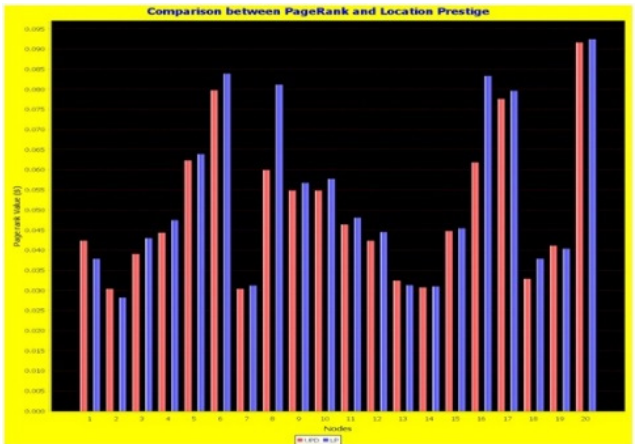

Figure 4. Comparison of UPD and PPS values.

4. Conclusion

This work uses the prestige score of each webpage using

the link structure. It determines the most significant and the least significant page. It is proposed the proximity prestige method that has higher performance. Necessary changes are done in the existing algorithm for finding the rank values. Incremental iteration is used in computing the rank value of the page. This reduces the computational cost in terms of time. And convergence occurs at a faster rate. In this work iteration process converges to a determined point and each page has a fixed rank value. The working scenario of uniform probability distribution and proximity work is exhibited using sample datasets and real dataset within a sub domain to determine the rank value. Most of the results prove that the proposed method outperformed existing pagerank algorithm by increase in the rank value thus increasing the chance of being a more relevant web page. As a future work this can be analysed for the performance with the other web structure algorithms and can be implemented through map reduce framework in order to increase the performance.

5. References

1. Brin S, Page L. The anatomy of a large scale hyper textual web search engine. Computer Network and ISDN Systems. 1998 Apr; 30(1–7):107–17.
2. Kim SJ, Lee SH. An improved computation of the page rank algorithm. Proceedings of the 24th BCS-IRSG European Colloquium on IR Research: Advances in Information Retrieval; 2002. p. 73–85.
3. Liu B. Web data mining, exploring hyperlinks, contents, and usage data. Springer-Verlag New York; 2006.
4. Data mining the web uncovering patterns in web content. Structure, and Usage [Internet]. [cited 2006 Jul 17]. Available from: http://onlinelibrary.wiley.com/book/10.1002/0470108096.
5. Scime A. Web mining applications and techniques. IGI Publishing; 2004 Aug.
6. Chakrabarti S. Mining the web - discovering knowledge from hypertext data. Morgan Kaufmann, 1st (Edn); 2002 Oct.
7. Kim KS, Choi YS. Incremental iteration method for fast page rank computation. Proceedings of the 9th International Conference on Ubiquitous Information Management and Communication; 2015.
8. Manning C, Raghavan DP, Schütze H. An introduction to information retrieval. Cambridge University Press; 2008 Jul.
9. Praba VL, Vasantha T. Efficient hyperlink analysis using robust proportionate prestige score in page rank algorithm. Applied Soft Computing. 2014 Nov; 24:86–94.
10. Kumar NKS, Kumar KK, Rajkumar N, Amsavalli K. Search engine optimization by fuzzy classification and prediction. Indian Journal of Science and Technology. 2016 Jan; 9(2):1–5.
11. Rani SG, Mageswari MAS. Link-click-concept based ranking algorithm for ranking search results. Indian Journal of Science and Technology. 2014 Oct; 7(10):1712–19.