Discussion on Damping Factor Value in PageRank Computation

Article *in* International Journal of Intelligent Systems and Applications · September 2017

DOI: 10.5815/ijisa.2017.09.03

3 authors, including:

Atul Kumar Srivastava
Banaras Hindu University
10 PUBLICATIONS **13** CITATIONS

I.J. Intelligent Systems and Applications, **2017, 9, 19-28**
Published Online September 2017 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijisa.2017.09.03

Discussion on Damping Factor Value in PageRank Computation

Atul Kumar Srivastava
Department of Computer Science, Institute of Science, Banaras Hindu University, Varanasi -221005, India
E-mail: atulbhuphd@gmail.com

Rakhi Garg
Department of Computer Science, MMV, Banaras Hindu University, Varanasi -221005, India
E-mail: rgarg@bhu.ac.in

P. K. Mishra
Department of Computer Science, Institute of Science, Banaras Hindu University, Varanasi -221005, India
E-mail: mishra@bhu.ac.in

Received: 17 January 2017; Accepted: 10 April 2017; Published: 08 September 2017

***Abstract*—**Web search engines use various ranking methods to determine the order of web pages displayed on the Search Engine Result Page (SERP). PageRank is one of the popular and widely used ranking method. PageRank of any web page can be defined as a fraction of time a random web surfer spends on that web page on average. The PageRank method is a stationary distribution of a stochastic method whose states are web pages of the Web graph. This stochastic method is acquired by combining the hyperlink matrix of the web graph and a trivial uniform process. This combination is needed to make primitive so that stationary distribution is well defined. The combination depends on the value of damping factor $\alpha \in [0,1]$ in the computation of PageRank. The damping factor parameter state that how much time random web surfer follow hyperlink structure than teleporting. The value of α is exceptionally empirical and in current scenario α = 0.85 is considered as suggested by Brin and Page. If we take α =0.8 then we can say that out of total time, 80% of time is taken by the random web surfer to follow the hyperlink structure and 20% time they teleport to new web pages randomly. Today web surfer gets worn out too early on the web because of non-availability of relevant information and they can easily teleport to new web pages rather than following hyperlink structure. So we have to choose some value of damping factor other than 0.85. In this paper, we have given an experimental analysis of PageRank computation for different value of the damping factor. We have observed that for value of α=0.7, PageRank method takes fewer numbers of iterations to converge than α=0.85, and for these values of α the top 25 web pages returned by PageRank method in the SERP are almost same, only some of them exchange their positions. From the experimental results it is observed that value of damping factor α=0.7 takes approximate 25-30% fewer numbers of iterations than α=0.85 to get closely identical web pages in top 25 result pages for personalized web search, selective crawling, intra-web search engine.

***Index Terms*—**PageRank method, Power method, Web matrix, Markov model, Eigenvalue problem.

I. Introduction

Due to use of Internet to a greater extent, web search engines have become the most important Internet tools to retrieve relevant information. Currently, thousands of web search engines have emerge in recent years based on various ranking method [1]. Google has become one of the most popular web search engine due to its ranking method called PageRank. PageRank computation is Static computation *i.e.* it is computed offline for every web page and independent of search query [2, 3, 4]. These computation nature makes it popular, and it is used in many applications like inverted index reordering, selective crawling, ranking sports team, clustering of similar web pages, Bioinformatics, Network analysis, Website search engine [5, 6]. PageRank is a hyperlink structure based method that computes the rank of all web pages indexed in the Internet by the Google's web crawler. To find PageRank one has to compute stable distribution of hyperlink matrix which is based on the web graph structure [7]. Since web contains huge amount of data so it is important that this method must be fast and needs to be accurate and efficient as much as possible [8, 9].

Langville, Brin et al, defined PageRank as the stationary distribution of a stochastic method whose states are the nodes of Web graph [7, 10]. According to them PageRank can be stated as follows: Let a random web user starts surfing from an arbitrary web page, and at every time they navigate to next web page by selecting one of the hyperlinks from the current web page. In initial approximation, we could define *PageRank as the ratio of*

time random web surfer spent on that web page to the number of web pages visited on average. However, according to some researcher's, this definition would not be appropriate for certain types of web pages that form a loop. Web pages in the loop may point to each other but do not point to any web page outside loop. These web pages only accumulate rank and do not distribute rank to other web pages and this property is called *rank sink* [8, 11, 12, 13]. To resolve rank sink issue, at every step the random web user has the choice to choose any out-link with probability α, and will restart from another node of *n* web pages chosen at random with probability $(1-\alpha)$, where α is a damping factor [7, 14]. Brin and Page, the founders of Google web search engine, suggest to take the value of damping factor α =0.85. The reason of selection of value of α=0.85 is not clear that motivates us to perform experiment to find out the reason of dependency of α on the efficiency of PageRank. In this paper, we discuss experimental effect of damping factor on PageRank computation and suggest new value of damping factor other than 0.85 as it gives better results.

II. Motivation and Literature Survey

The recent survey estimated that web is the largest collection of data distributed over approximate 200 million web servers containing greater than 15 billion web pages, and larger than 600 million hosts [15, 16]. Due to dynamically growing nature of web and needs of user's query the role of web page ranking method becomes very crucial in any information retrieval system. Today, whenever a user searches any query then search engines returns thousands or millions of results regarding query keyword. Web users do not have the much time and the patience to go through all returned pages to find the relevant one in which they are interested, and some analysis shows that the 70% web users' don't even goes beyond the first page of SERP [17]. Therefore, it is very important for ranking method to keep the desired result within top few web pages, otherwise web search engine could be thought as improper. The PageRank method highly depends on the value of damping factor $\alpha \in [0,1]$ that plays a major role in the convergence of PageRank computation [18, 19]. However, variation in value of α not only affect rank value of web pages, but also change their displayed order in SERP [18, 19, 20]. The number of iterations taken by PageRank method grows on increasing the value of α and required more numerical precision to converge as value of $\alpha \to 1$. PageRank of the web graph lies between a true uniform distribution ($\alpha = 1$) and a meaningless artificial teleportation distribution, mostly of irrelevant web pages ($\alpha = 0$). It is easy to observe that picking small value of α is unsuitable, because too much weight provided to the identical artificial teleportation matrix [19, 21]. In this paper, we discuss the effect of different value of damping factor on computation of PageRank on various datasets.

Few researchers have discussed about effect of damping factor on PageRank computation. Chritopher Engstorm et al, observe convergence rate of PageRank computation on various damping factor value by changing the weight of some web pages in the Web graph and conclude that number of iteration taken to converge by PageRank algorithm increases as damping factor value goes closer to 1 [18]. Paolo Boldi et al, have done various analysis on PageRank computation on different damping factor value, and they observed that value of damping factor which is very closer to 1 provide true web structure and they also obtained the Maclaurin polynomial for PageRank that is independent from α and induces interesting rankings [12]. Langville et al, state that value of damping factor shows how much time random web user spent on surfing the hyperlink structure than teleporting the web [13, 22]. Guan et al, have done citation analysis of scientific research articles by using PageRank and used value of damping factor $\alpha = 0.5$ and state that this value of α gives more true ranking for analysis of scientific articles [23].

In earlier years user can easily get relevant information due to small size of data on web so the value of damping factor is taken as $\alpha = 0.85$ by Brin & Page, But nowadays there are more irrelevant information available on web so that random web user does not follow hyperlink structure and easily teleport to new web pages very frequently. In order to provide true link structure of web, we have to choose value of α that should be different from $\alpha = 0.85$. In some applications like intra-web search engine, selective crawling, traffic analysis, bibliography ranking. The value of Damping factor $\alpha = 0.7$ provides relevant ranking result same as for $\alpha = 0.85$ and take less numbers of iteration and time. In this paper, we have done experiment on various datasets and observed the result for different damping factor value and concludes that $\alpha = 0.7$ gives more true structure of web and take fewer numbers of iterations and time to converge the PageRank method than for $\alpha = 0.85$.

The rest of paper is divided into following sections. Section II describes some mathematical terminology used in the paper. Section III includes basics of PageRank method and its computation and structure of the real Web graph. In Section IV, we discuss Improved PageRank method and the importance of Power method in PageRank computation. In Section V, we have observed the effect of damping factor on ranking of web pages as well as the convergence of PageRank Power method. In Section VI we briefly discuss about implementation details and observations. Finally Section VII concludes the experimental result.

III. Mathematical Terminology

We treat the HTML "points-to" relation on web pages as a directed graph, and represents it by two sets: *G = (V, E)*, Where *V* is the set of nodes corresponding to web pages, and *E* represents set of edges corresponding to hyperlinks between web pages. Some common mathematical definition and notations which will be used in this paper are as follows: [4, 24]

- **Hyperlink Matrix:** It shows the surfing behavior

of random surfer. The elements of hyperlink matrix *M* are assigned by following Eq. (1):

$$M_{ij} = \begin{cases} \frac{1}{Outdegree\ of\ P_i}, if\ i \to j \\ 0, \qquad otherwise \end{cases} \quad (1)$$

- **Sparse Matrix*:*** A sparse matrix is a matrix in which most of the elements are zero. By contrast, if most of the elements are nonzero, then the matrix is considered dense.
- **Stochastic Matrix*:*** A stochastic matrix also called transition matrix, probability matrix or markov matrix used to describe transition of markov model or random surfer. It is a matrix whose entries in each row are non-negative real numbers and sum equal to one.
- **Aperiodic*:*** A state i is said to be periodic with period $k \geq 1$ if k is the smallest number such that all paths starting from state i back to state i have a length that is multiple of k. If $k = 1$ then state is not periodic. A markov model or Web graph is aperiodic if and only if all states or web pages are aperiodic.
- **Irreducible*:*** A Web graph is irreducible when it is strongly connected. A directed Web graph $G = (V, E)$ is strongly connected iff $\forall$ pair of nodes $(u, v) \in V$, there is a path from u to v.
- **Primitive matrix*:*** A transition matrix is primitive matrix if it is *irreducible* and *aperiodic*.
- **Dangling Node*:*** A node is said to be dangling node if it does not point to any other node. In hyperlink matrix, row corresponding to dangling node contains only zero.
- **PageRank vector** $(\boldsymbol{\pi})$***:*** It is a column vector which contains rank value corresponding to every web page of Web graph.
- **Dangling node vector *(a):*** It is a binary column vector that contains One corresponding to every dangling nodes and zero for other nodes.
- **e:** It is an identity column vector of $1 * n$.
- $\boldsymbol{\alpha}$***:*** It denotes Damping factor value
- $\boldsymbol{\epsilon}$ ***:*** Tolerance value used for convergence of PageRank vector

IV. Basic PageRank Model

The PageRank method proposed by Brin and Page in 1998, is used in Google Search engine as a basic method [7, 8]. It states that importance of web pages is determined by the number of hyperlinks pointing to it as well as rank of those web pages which are pointing to it. It is also measured as a probability that the random web surfer visits any particular web page.

Definition 1 *If web page P_j have I_j outlinks and one of these links pointing to another web page P_i, then web page P_j will pass $\frac{1}{I_j}$ i.e. $\frac{1}{Outdegree\ of\ P_j}$ importance or rank to web pageP_i. So, PageRank of pageP_i is the sum of all the contribution made by web pages pointing to it.*

Let O_{P_i} is the set of pages pointing to P_i then following Eq. (2) shows mathematical equation to compute PageRank:

$$rank_{k+1}P_i = \sum_{P_j \in O_{P_i}} \frac{rank_k P_j}{outdegree\ of\ P_j} \quad (2)$$

Where *k* denotes iteration number. In the beginning of iterative procedure rank of all web pages are unknown, so we assign rank of each web pages to $\frac{1}{n}$, where *n* is the total number of web pages. After that this iterative equation is successively repeated and substitute the value of $rank\ P_i$ in next iteration. This iterative equation begin with $rank_0 P_i = \frac{1}{n}$ for every pages and repeated until the PageRank score converged to some final stable values. By Eq. (2) we can compute PageRank value of web pages one at a time. Another way to compute PageRank by converting Eq. (2) in matrix format and compute PageRank vector of $1 * n$, which hold rank of each web pages. Let *M* be the hyperlink matrix of order $n \times n$ and π_k is the PageRank vector at k^{th} iteration then compute PageRank of web pages by using Power method by Eq. (3):

$$\pi_{k+1} = \pi_k M \quad (3)$$

Where *M* is a sparse matrix due to large number of dangling Web pages. Consider a random web surfer start surfing from any of *n* web pages with equal probability. Afterwards in first iteration initial vector π_0 initialize with $\frac{1}{n}$ for each web page. Subsequently after one step, the distribution vector of the random surfer will be $M\pi_0$. After second steps it will be $M^2\pi_0$, and so on. In general, we can state that, by iterative multiplying i^{th} times *M* to initial vector π_0 *i.e.* $M^i\pi_0$, will give us the distribution vector of random web user after i^{th} steps. This kind of nature is known as an example of Markov method or Markov matrix [14, 25]. It is clear that for any starting PageRank vector, the markov method converges to a unique stationary vector when matrix *M* is *stochastic* and *primitive i.e.* when the Web graph should be strongly connected and there should be no dangling web page in the Web graph [26, 27]. These two characteristics of Web graph ensures the existence of unique PageRank vector. However, it is impossible in practical scenario. Earlier, Researchers have observed the structure of Web as shown in Fig. 1 that contains basically three main components: *SCC, IN-Component, Out-Component* [28, 29].

First one is *strongly connected component (SCC)* which is the largest component in terms of size. Second, one is *IN-component* that contains the web pages that have outgoing links to SCC, but could not have incoming links from the SCC and the third one is *Out-Component*, consisting of web pages having incoming links from SCC but do not have outgoing links to reach SCC. There are

other three components called *Tendrils, Tubes and Isolated components* which also have their importance in the Web graph structure. *Tendrils* are mainly two types: One type of tendrils contains the web pages which are reachable from the IN-Component but are not having link to reach the IN-Component.

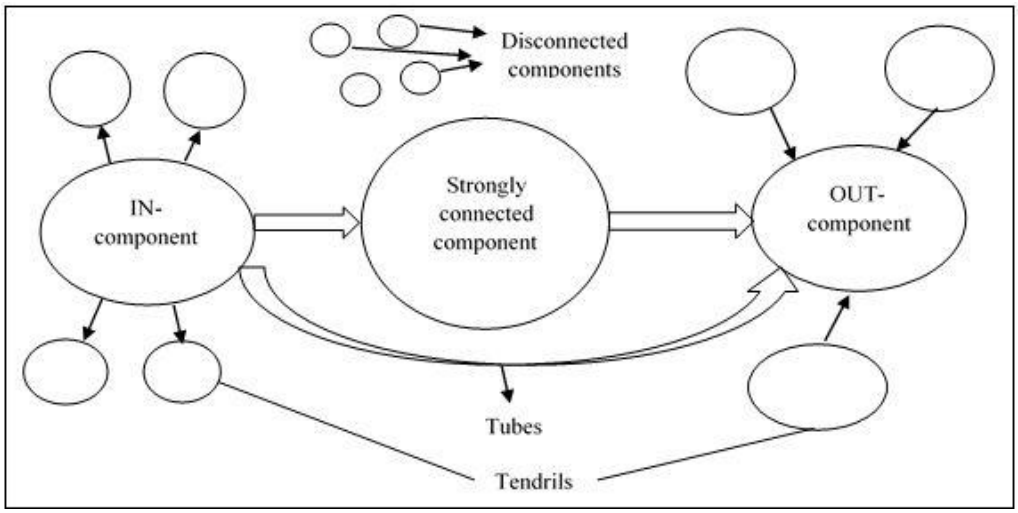


Fig.1. Structure of Web

Second type of tendrils can reach the OUT-Component, but not able to reach from the OUT-Component. *Tubes* contains small number of web pages that are reachable from the IN-Component and able to reach the OUT-Component, but it does not able to reach the SCC or to be reached from the SCC. *Isolated components* are those that are not able to reachable from the large components like SCC, IN, OUT- components, and not able to reach those components. Among these structures tendrils, OUT-Component and Isolated Component violates the rules needed for the Markov iteration method to converge to a tolerance limit value.

Consider a random web surfer that enters into OUT-Component, after that they can never leave from this component because the web pages does not have out-links to any other components. As a result, then random web user beginning in either the SCC or IN-Component are going to set-up in either the OUT component or a tendril off the IN-Component. Therefore, no web page in the SCC or IN-component winds up with any probability of a web user being there. If we take this probability as a component for measuring the importance of web page, then we determine incorrectly that nothing in the SCC or IN-Component is of any importance. Therefore, PageRank is generally improved to avoid such anomalies. There are two problems that should be avoided. First is the dangling nodes, which is a page that has no outgoing links and second is Rank Sink issue, a group of web pages that linked with each other but none of them link to any other pages.

V. Improvement of PageRank Considering Dangling Node & Rank Sink Issue

Brin and Page have modified basic PageRank method in order to resolve above two issues [8, 30]. To solve the dangling node problem, they have used the property of stochastic matrix. They have converted hyperlink matrix *M* to stochastic matrix *S* by replacing all rows having *"0"* entries for every web page with $\frac{1}{n}e^T$. After this adjustment, any of *n* pages at random are visited once user reach to a dangling node [13]. This adjustment can be represented mathematically by Eq. (4):

$$S = M + a(\frac{1}{n}ee^T) \tag{4}$$

One of the problem after this adjustment is that it can't guaranteed the convergence of the PageRank method [13, 31]. To overcome with this issue just make matrix *S* primitive [8, 13]. Generally, web user follow the hyperlink structure of web, at some times when they get bored then stop to follow the hyperlink structure and teleport to new web page with probability α called damping factor, by entering a new web page address in the browser's URL. When this happened, they begin surfing again, until the next teleport required and so on. So to make primitive adjustment of matrix *S*, formulate a new matrix *H* denoted by Eq. (5) [13]:

$$H = \alpha S + (1-\alpha)\frac{1}{n}ee^T \tag{5}$$

Where α denotes the damping factor and takes value $0 \leq \alpha \leq 1$.

The damping factor parameter states that how much time random web surfer follow hyperlink structures than teleporting. There are several properties of matrix *H*: Matrix *H* is stochastic, irreducible, aperiodic and primitive, and completely dense and huge, which has a bad impact on the computation of PageRank [32, 33]. Eq. (5) is a linear PageRank equation that is solved by either direct methods or iterative methods. However, direct method does not perform very well for sparse matrix as compared to iterative methods. So PageRank method

solved by Power method that is one of the simplest iterative method to find the dominant eigenvalue and eigenvector of a matrix is used [4, 7, 25]. PageRank can be implemented by Power method mathematically as stated below:

$$\pi_{k+1} = \pi_k H$$
$$\pi_{k+1} = \pi_k [\alpha S + \frac{(1-\alpha)}{n} e e^T]$$
$$\pi_{k+1} = \pi_k [\alpha (M + \frac{1}{n} a e^T) + \frac{(1-\alpha)}{n} e e^T]$$
$$\pi_{k+1} = \alpha \pi_k M + [\alpha \pi_k a + \frac{(1-\alpha)}{n} e^T] \quad (6)$$

From Eq. (6), it is clear that there is no need to store matrix S and H, only rank one component of a vector and e are needed. We know that vector matrix multiplication is $O(n)$ because it contains approximation *10 nonzero per row* [34]. The Power method is slowest among other iterative methods like Gauss-Seidel, Jacobi, restarted GMRES etc., but it has advantage over these methods *i.e.* it is a matrix free iterative method so no matrix computation is required and because of this, it is preferred while storage of the hyperlink matrix is very large.

Brin and Page, Amy N. Langville and Carl D. Meyer, Pretto, L discussed in their paper and also state that Power method needs only 50-100 iteration to converge the algorithm and as we know that each iteration of the Power method requires $O(n)$ computation so it can take approximate $50\ O(n)$ to $100\ O(n)$ Power iteration [7, 13].

VI. Effect of Damping Factor on PageRank Algorithm

As it is stated above that damping factor α shows the proportion of time the random web surfer follows the hyperlinks contrary to teleporting. So damping factor α plays an important role in PageRank computation by Power method shown in Eq. (7) [8, 25]:

$$\pi_{k+1} = \pi_k [\alpha S + \frac{(1-\alpha)}{n} E] \quad (7)$$

Where $E = ee^T$ is the teleportation matrix of $n \times n$, α controls priority given to the hyperlink structure over artificial teleportation matrix E. Many researchers state that the value of damping factor α affect the convergence rate of PageRank Power method [11, 12, 14, 18, 19]. We have performed an experiment to observe the impact of damping factor on the convergence rate of PageRank algorithm, and ordering of web pages displayed on web search engine on various datasets by using following Eq. (8) & algorithm (1):

$$\pi_{k+1}^i = \alpha \left(\sum_{\forall j \to i} \pi_k^j O^j \right) + \frac{(1-\alpha)}{n} + \frac{\alpha}{n} \left(\sum_{\forall m \text{ i.e. } dangling\ nodes} \pi_k^m \right) \quad (8)$$

Where π_{k+1}^i denotes PageRank vector at k^{th} iteration, O^j represents out-degree of web-page j.

Algorithm 1 PageRank algorithm by using Power method

procedure $\pi_0 = PowerMethod()$
  $\pi_0 = \frac{1}{n}$ (a row vector)
  H = row − normalized hyperlink matrix
  α = Damping factor value
  ϵ = Convergence tolerance value
  n = Number of Web pages in Web graph
  $k = 1$
  do
    $\pi_{k+1} = H\ \pi_k$
    $\delta = ||\pi_{k+1} - \pi_k||$
    $k = k + 1$
  while $(\delta < \epsilon)$
end procedure

VII. Experimental Analysis and Discussion

In this section, we have described experiments performed on various real datasets taken from *cs.toronto.edu* and *snap.stanford.edu* website. These websites contain many data sets of distinct queries [35, 36]. Subsection 1 briefly describe the data and the experimental setup and subsection 2 observe the performance of PageRank on the different value of damping factor α varies between 0 and 1.

A. Data and Experimental setup

We have implemented PageRank Power method in *JAVA (JDK 1.8)* language. We have done experiment on single Linux machine *(Ubuntu 14.04 LTS), Intel Core i5 CPU 3.2 GHz, 4 GB RAM* on following datasets:

Table 1. Datasets with their attributes

Dataset	No. of Web pages	No. of Dangling Web pages
YouTube dataset (D1)	1157827	783042(0.67%)
Road network dataset (D2)	1971281	6075 (0.30%)
Wikitalk dataset (D3)	2394385	2246783 (93%)

First dataset, YouTube is a video-sharing web site that contains a social web network. YouTube users can create various social groups to connect and interact with other YouTube users. Here YouTube graph represented by $G:(V,E)$ *Where V is node & E edge of Web graph represented YouTube users & links between various users respectively.* Second Dataset, A road network of California, where Intersections and endpoints of roads represents nodes and the roads connecting these intersections or road endpoints represents directed edges. Third Dataset, Wikipedia dataset shows that every registered user has a talk page, which he and other users can edit in order to communicate and discuss updates to various articles on Wikipedia. Nodes in the network represented by Wikipedia users and a directed edge from node i to node j represented by that user i at least once edited a talk page of user j.

Due to sparse matrix storage we stored these web graphs into Hash-map data structure for the computation of PageRank. Since in the PageRank computation, only nonzero entry of the hyperlink matrix is required, so we store only the non-zero entry in Hash-map data-structure that reduce not only storage but it also provides fast access of data [2, 10, 27]. For large datasets Hash-map data structure would be better in terms of accessing of the element and storage of elements than hyperlink matrix. We have implemented Hash-map data-structure in Java language using Guava library provided by Google [37].

B. Observation of damping factor on PageRank computation

We have applied PageRank method on each dataset till convergence and observed the convergence speed for different damping factor value. Following Fig. 2 shows the number of iteration taken to converge PageRank method on various damping factor. The plotted graph depicts that as the value of damping factor increases, the number of iteration taken by PageRank method to converge also increases for tolerance value $\epsilon = 10^{-7}$. From Fig. 2, we can say that there is slight change in the number of iteration when $0 < \alpha < 0.65$ but graph curve increases rapidly when value of $\alpha > 0.65$ and tends to *1*.

C. Damping factor value versus Convergence Rate:

Earlier researchers have observed that value of damping factor not only controls the convergence of PageRank method but also affects the ordering of web pages returned by the PageRank method [18]. As we can see that in Fig. 2, that $\alpha = 0.7$ take only 12 iterations to converge when $\epsilon = 10^{-7}$ for 1157827 nodes. As $\alpha \to 1$ this number becomes prohibitive and for the large dataset, the choice of $\alpha \to 0.85$ still requires too much time to converge. Earlier *Brin and Page* chosen value of $\alpha = 0.85$ because the large value of α gives more weight to the true link structure of web while smaller values of α

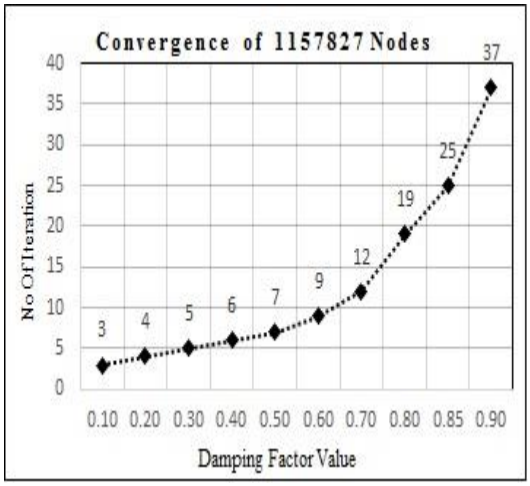


Fig.2(a)

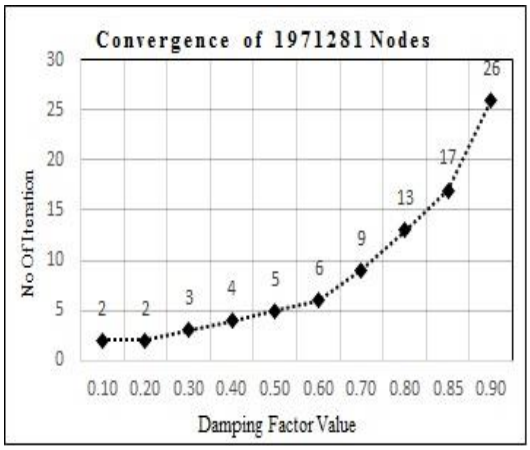


Fig.2(b)

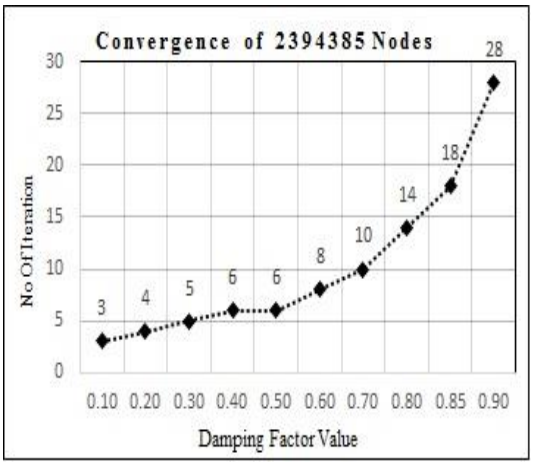


Fig.2(c)

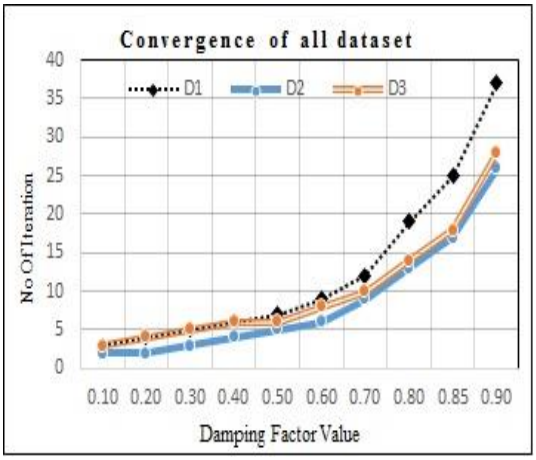


Fig.2(d)

Fig.2. Number of iteration taken to converge for the value of $\alpha = (0,1)$ on tolerance value $\epsilon = 10^{-7}$. (a) For 1157827 nodes, (b) for 1971281 nodes, (c) for 2394385 nodes, (d) for all data nodes together.

increases the influence of the artificial probability vector. However, nowadays random web user does not follow hyperlink structure and teleport frequently to other web pages due to a huge amount of irrelevant data on the web. So the value of damping factor between $\alpha \in (0.7, 0.8)$ gives truer behavior of hyperlink structure, and it also takes fewer numbers of iterations to converge the method. Damping factor value $\alpha = 0.7$ would give more weight to artificial teleportation vector that would be useful in personalized PageRank, Domain based search engine and many other application where artificial teleportation vector plays a major role.

Fig. 2 shows the number of iteration taken by PageRank method to converge on various dataset with tolerance value $\epsilon = 10^{-7}$. Number of iteration taken to converge on PageRank method does not totally depend on data size, it also depends on several properties of dataset *i.e.* the number of dangling nodes, connectivity of nodes. As we can see that from Fig. 2(d) and Table 1 for dataset *D3* PageRank method takes fewer numbers of iterations to converge than *D1*.

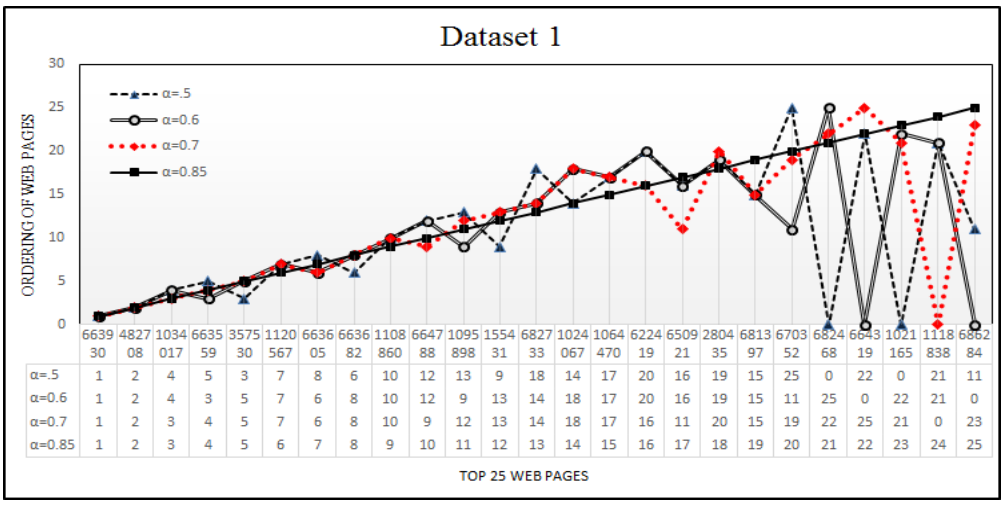
In this experiment we have taken, only top 25 web pages return by PageRank method because SERP (Search engine result pages) contains almost 20 to 30 web pages in their first result page that are more relevant to the web user, and on average 90% users do not go beyond the first result page for any particular query [9, 11]. If any web page indexed in the top 25 position for $\alpha = 0.85$ ($\alpha_{0.85}$) but cannot get position in top 25 for $\alpha = 0.7$ ($\alpha_{0.7}$), then we have taken a parameter λ to show that whether page is indexed in top 25 or not, parameter λ is defined as follows:

$$\lambda = \begin{cases} 0, & if\ web\ page\ is\ not\ get\ position\ in\ top\ 25 \\ 1 \le \lambda \le 25, & when\ it\ is\ indexed\ in\ top\ 25 \end{cases} \quad (9)$$

As example in Fig. 3(a) web page *1454* got *21st* position for $\alpha_{0.85}$ but it can't got position in top 25 for $\alpha_{0.7}$ so we give it $\lambda = 0$. In Fig. 3 Straight line shows the ordering of web pages for damping factor $\alpha_{0.85}$. In Fig 3(a, b, c) we can see that web pages for $\alpha_{0.7}$ got almost same ordering *w.r.t.* $\alpha_{0.85}$.

In Fig. 3a) only one web page could not index in top 25 *i.e. Page No. 1964*, In Fig. 3b) Three web page could not index in top 25 *i.e. Page No. 1259079, 737950, 749665* and In Fig. 3c) only one web page could not index in top 25 *i.e. Page No. 1118838* for damping factor $\alpha_{0.7}$ *w.r.t.* $\alpha_{0.85}$ on *D1, D2, D3* dataset respectively.

We compare the top 25 web pages returned by $\alpha_{0.85}$ with other damping factor α value in terms of two useful and informative parameters, namely *Relative Common Web pages (RCWP):* It Indicates the number of web pages got the position in top 25 for both damping factor value. Second one is *Web Pages got a different index w.r.t.* $\alpha_{0.85}$ (N_i): It Shows the number of web pages got a different index/order number for both damping factor value.



	6639 30	4827 08	1034 017	6635 59	3575 30	1120 567	6636 05	6636 82	1108 860	6647 88	1095 898	1554 31	6827 33	1024 067	1064 470	6224 19	6509 21	2804 35	6813 97	6703 52	6824 68	6643 19	1021 165	1118 838	6862 84
α=.5	1	2	4	5	3	7	8	6	10	12	13	9	18	14	17	20	16	19	15	25	0	22	0	21	11
α=0.6	1	2	4	3	5	7	6	8	10	12	9	13	14	18	17	20	16	19	15	11	25	0	22	21	0
α=0.7	1	2	3	4	5	7	6	8	10	9	12	13	14	18	17	16	11	20	15	19	22	25	21	0	23
α=0.85	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25

Fig.3(a)

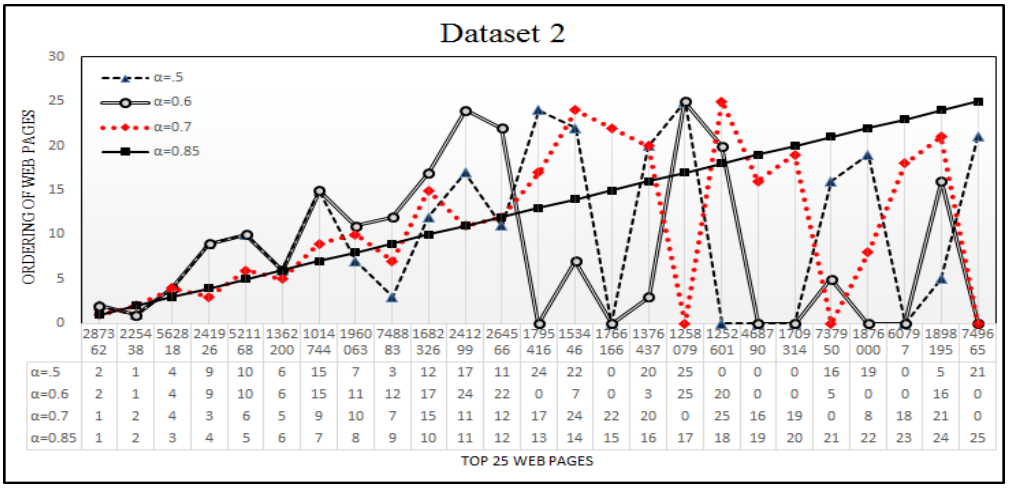


Fig.3(b)

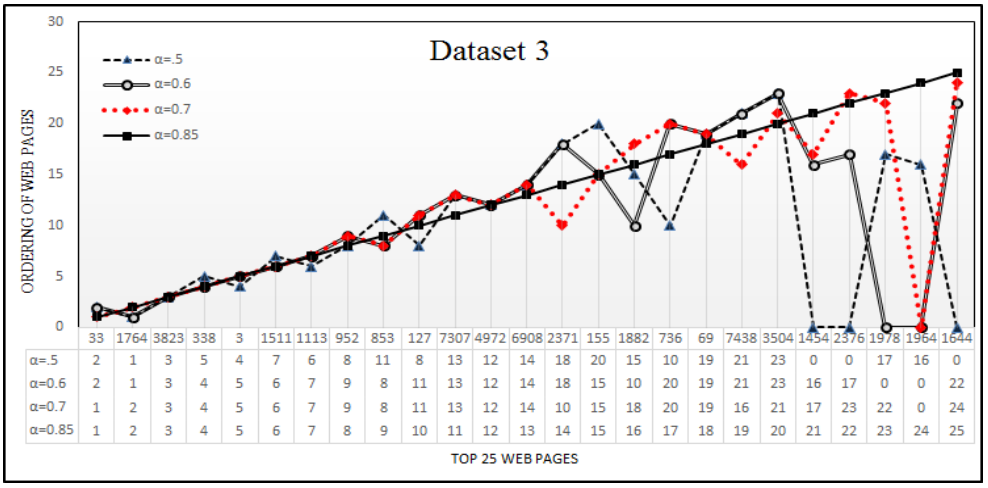


Fig.3(c)

Fig.3. This figure a, b and c, shows relative comparison of ordering number of top 25 web pages for different damping factor on *D*1, *D*2, *D*3 dataset respectively.

We define the following notation and expressions used to compute these parameters:

- WP_α: Top 25 web pages returned by PageRank method for damping factor value α.
- $RCWP_\alpha$: Number of common web pages for damping factor value α with respect to $\alpha_{0.85}$.
- N_α : Number of web pages got the different position for damping factor value α *w.r.t.* $\alpha_{0.85}$.
- WP_α^i: Web pages on index i for damping factor value α.

Relative Common Web Pages (RCWP) computed by using following Eq. (10):

$$RCWP_{any\ value\ from\ 0.7,0.6,0.5} = WP_{0.85} \cap WP_{any\ value\ from\ 0.7,0.6,0.5} \quad (10)$$

More relative common web pages $RCWP_\alpha$ represent that both damping factor value indexes almost same web pages in top 25. So we can use the value of damping factor $\alpha_{0.7}$ in-place of $\alpha_{0.85}$ for certain types of web pages as discussed in above section 5.2.1. Second Parameter *Web Pages got a different index w.r.t.* $\alpha_{0.85}$ (N_i) is computed by following Eq. (11):

$$N_{any\ value\ from\ 0.7,0.6,0.5} = \sum_{1 \le i \le 25} (WP_{0.85}^i \neq WP_{any\ value\ from\ 0.7,0.6,0.5}^i) \quad (11)$$

Table 2. Percentage of *RCWP*, N_i & Number of iteration taken by PageRank method

	Relative Common Web Pages (RCWP)			**Number of Web Pages Got different position (*Ni*)**			**Number of Iteration (I_α)**			
	$CWP_{0.7}$	$CWP_{0.6}$	$CWP_{0.5}$	$N_{0.7}$	$N_{0.6}$	$N_{0.5}$	$I_{0.85}$	$I_{0.7}$	$I_{0.6}$	$I_{0.5}$
YouTube	96%	92%	88%	64%	72%	88%	25	12	09	07
Road n/w	88%	72%	80%	84%	100%	100%	17	09	06	05
Wikitalk	96%	92%	92%	72%	84%	84%	18	10	08	06

VIII. Conclusion

Brin and Page have taken the value of damping factor $\alpha_{0.85}$ which is still used by many researchers. Langville et.al, state that α controls the proportion of time the random web user follows the hyperlink structure contrary to teleporting. Nowadays, there is a lot of information on Web, which is not significant to users, so that random web user does not follow hyperlink structure for long time, and they are willing to teleport to any other web page more often. Under such condition, there must be a change in value of α to get the relevant result in less time. Many researchers observed that damping factor controls the convergence speed of PageRank algorithm. We have observed in the experiment that PageRank method takes more iterations and time to converge the algorithm for $\alpha_{0.85}$ rather than $\alpha_{0.7}$, and the returned result of web pages are almost same, only their ordering are changed. By performing the experiment on different datasets we conclude that the number of relevant web pages generated by $\alpha_{0.7}$ & $\alpha_{0.85}$ remains same but only difference is that computing time is less in case of $\alpha_{0.7}$. We can use this damping factor value to speed up the convergence of PageRank algorithm in many systems like web-site search engine, domain based search engines, clustering of web pages, network analysis. Where minor difference in ordering of web pages does not play an important role. In future, we will perform more experiments regarding the value of damping factor and other factors, which can speed up PageRank computation efficiently.

References

[1] Sun, H. and Wei, Y., (2006) "A note on the PageRank algorithm" Applied Mathematics and computation, Vol. 179 No.2, pp.799-806.

[2] Bianchini, M., Gori, M. and Scarselli, F., (2005) "Inside pagerank" ACM Transactions on Internet Technology (TOIT), Vol. 5 No. 1, pp.92-128.

[3] F. Crestani, M. Girolami, and C. J. van Rijsbergen (2002) "Advances in Information Retrieval" number 2291 in LNCS, pp. 73–85.

[4] Rajaraman, A. and Ullman, J.D., (2012) "Mining of massive datasets" Cambridge: Cambridge University Press, Vol. 1.

[5] Muhammad Mahbubur Rah-man,"Mining Social Data to Extract Intellectual Knowledge", International Journal of Intelligent Systems and Applica-tions(IJISA), vol.4, no.10, pp.15-24, 2012. DOI: 10.5815/ijisa.2012.10.02.

[6] Henzinger, M. R. (2001) "Hyperlink analysis for the web", Internet Computing, IEEE, Vol.5 Issue.1, pp. 45-50.

[7] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd (1999) "The PageRank citation ranking: Bringing order to the web" Technical Report 66, Stanford University, Available at http://dbpubs.stanford.edu/pub/1999-66.

[8] Brin, S., and Page, L. (2012) "Reprint of: The anatomy of a large-scale hyper-textual web search engine" Computer networks, Vol. 56 No.18, pp. 3825-3833.

[9] Sargolzaei, P. and Soleymani, F., (2010) "PageRank problem, survey and future research directions" In International Mathematical Forum, Vol. 5 No. 19, pp. 937-956.

[10] Langville, A. N., and Meyer, C. D. (2004) "Deeper inside pagerank" Internet Mathematics, Vol.1 No.3, pp. 335-380.

[11] Avrachenkov, K., Litvak, N., and Pham, K. S. (2008) "A singular perturbation approach for choosing the PageRank damping factor" Internet Mathematics}, Vol.5 No. (1-2), pp. 47-69.

[12] Boldi, P., Santini, M., and Vigna, S. (2005) "PageRank as a function of the damping factor" In Proceedings of the 14th international conference on World Wide Web, ACM, pp. 557-566.

[13] Langville, A.N., Meyer, C.D. (2006) "Google's PageRank and Beyond: The Science of Search Engine Rankings" Princeton University Press, Princeton.

[14] Brinkmeier, M. (2006) "PageRank revisited" ACM Transactions on Internet Technology (TOIT), Vol. 6 No. 3, pp. 282-301.

[15] Arasu, A., Novak, J., Tomkins, A., and Tomlin, J. (2002) "PageRank computation and the structure of the web: Experiments and algorithms" In Proceedings of the Eleventh International World Wide Web Conference, Poster Track pp. 107-117.

[16] Avrachenkov, K., and Litvak, N. (2006) "The effect of new links on Google PageRank" Stochastic Models, Vol. 22 No. 2, pp. 319-331.

[17] Borodin, A., Roberts, G. O., Rosenthal, J. S., and Tsaparas, P. (2005) "Link analysis ranking: algorithms, theory, and experiments" ACM Transactions on Internet Technology (TOIT), Vol. 5 No. 1, pp. 231-297.

[18] Baeza-Yates, R., Boldi, P., and Castillo, C. (2006) "Generalizing pagerank: Damping functions for link-based ranking algorithms" In Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval ACM, pp. 308-315.

[19] Bressan, M., and Peserico, E. (2010) "Choose the damping, choose the ranking" Journal of Discrete Algorithms, Vol. 8 No. 2, pp. 199-213.

[20] Melucci, M., and Pretto, L. (2007) "PageRank: When order changes" Springer Berlin Heidelberg, pp. 581-588.

[21] Engström, C. and Silvestrov, S., (2014) "Generalisation of the Damping Factor in PageRank for Weighted Networks" In Modern Problems in Insurance Mathematics, Springer International Publishing, pp. 313-333.

[22] Pretto, L. (2002) "A theoretical analysis of Google's PageRank" In String Processing and Information Retrieval, Springer Berlin Heidelberg, pp. 131-144.

[23] Ma, N., Guan, J. and Zhao, Y., (2008) "Bringing PageRank to the citation analysis" Information Processing & Management, Vol. 44 No.2, pp.800-810.

[24] Liu, B., (2007) "Web data mining: exploring hyperlinks, contents, and usage data" Springer Science & Business Media.

[25] Berkhin, Pavel (2005) "A survey on pagerank computing" Internet Mathematics, Vol. 2 No. 1 pp. 73-120.

[26] Berry, M. W., Drmac, Z., and Jessup, E. R. (1999) "Matrices, vector spaces, and information retrieval" SIAM review, Vol. 41 No. 2, pp. 335-362.

[27] Donato, D., Laura, L., Leonardi, S., and Millozzi, S. (2004) "Large scale properties of the web graph" The European Physical Journal B-Condensed Matter and Complex Systems, Vol. 38 No. 2, pp. 239-243.

[28] Kamvar, S., Haveliwala, T., Manning, C., and Golub, G. (2003) "Exploiting the block structure of the web for computing pagerank" Stanford University Technical Report.

[29] Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., and Wiener, J. L. (2000) "Graph structure in the web" Computer networks, Vol. 33 No. (1-6), pp. 309-320.
[30] Kim, S. J., & Lee, S. H. (2002) "An improved computation of the pagerank algorithm, Advances in Information Retrieval, Springer Berlin Heidelberg, pp. 73-85.
[31] Lempel, Ronny, and Shlomo Moran (2001) "SALSA: the stochastic approach for link-structure analysis" ACM Transactions on Information Systems (TOIS), Vol. 19 No.2, pp. 131-160.
[32] Gleich, David Francis, and Michael Saunders (2009) "Models and algorithms for pagerank sensitivity" Stanford University, 2009.
[33] Serra Capizzano, S. (2007) "Google PageRanking problem: The model and the analysis" In Dagstuhl Seminar Proceedings, Schloss Dagstuhl-Leibniz-Zentrum für Informatik.
[34] Varga, Richard S. (2009) "Matrix iterative analysis" Springer Science and Business Media, Vol. 27.
[35] Datasets for Experiments on Link Analysis Ranking Algorithm shttp://www.cs.toronto.edu/experiments/datasets/index.html
[36] Jure Leskovec and Andrej Krevl (2014) "Stanford Large Network Dataset Collection" http://snap.stanford.edu/data.
[37] "Guava: Google Core Libraries for Java" https://github.com/google/guava.

Authors' Profiles

Atul Kumar Srivastava received his Master of Computer Application (M.C.A.) degree from Banaras Hindu University, Varanasi in 2011. Currently he is pursuing Ph. D. in Department of Computer Science, Institute of Science, Banaras Hindu University. His research interests include Web mining and PageRank Optimization.

Rakhi Garg received her M. Sc. and Ph. D. degrees from Department of Computer Science, Banaras Hindu University, Varanasi, India in 1997 and 2012 respectively. She has more than 15 years of teaching experience at various Institutions. Since 2007, She has been working as a faculty member and In-charge in Department of Computer Science, Mahila Maha Vidayalya, Banaras Hindu University, Varanasi, India. Her research interests include Data Mining, Web Mining, Parallel and Distributed data mining.

P. K. Mishra is a Professor, in Department of Computer Science, Banaras Hindu University, India. He is also a Principal Investigator of the research projects at DST Centre for Interdisciplinary Mathematical Sciences, Banaras Hindu University. His research interests include Parallel and Distributed Computation, Computational Complexity, Parallel and Clustered Data Mining, High Performance Computing and VLSI Algorithms. Prof. Mishra has more than 70 publications in conference proceedings and journals. He is a reviewer and editor of various journals and senior member of IEEE.

How to cite this paper: Atul Kumar Srivastava, Rakhi Garg, P. K. Mishra, " Discussion on Damping Factor Value in PageRank Computation", International Journal of Intelligent Systems and Applications(IJISA), Vol.9, No.9, pp.19-28, 2017. DOI: 10.5815/ijisa.2017.09.03