Analysis of Successive Over Relaxation Method in PageRank Computation

Atul Kumar Srivastava[1], Mitali Srivastava[2], Rakhi Garg[3], P. K. Mishra[4]

[1, 2, 4] Department of Computer Science, Faculty of Science, Banaras Hindu University, Varanasi, India

[3]Computer Science Section, Mahila Maha Vidayalaya, Banaras Hindu University, Varanasi, India

Abstract: PageRank is one of the basic metric to compute the ranks of web pages used by Google search engine. PageRank value can be understood as a frequency of visiting any web page by random web user and thus it is interpreted as the measure of importance of the web page. Google search engines has used Power method to compute PageRank value which takes several weeks to converge the method due to huge size of web. Many researcher have proposed some other iterative methods like: *Gauss-Seidel, Successive Over Relaxation, Monte Carlo iterative method etc*. that takes less number of iterations to converge than Power method. In this paper, we have discussed Successive Over Relaxation method which is the improvement of Gauss-Seidel method to compute PageRank implemented on various datasets. On analyzing the result obtained after the execution of *Successive Over Relaxation* and *Gauss-Seidel* method to compute PageRank for a given dataset, we can say that the *Successive Over Relaxation* is better than *Gauss-Seidel* in terms of number of iterations required to converge.

Keywords: PageRank method, Successive Under Relaxation method, Successive Over Relaxation method.

Introduction

PageRank method is one of most popular and best-known method to compute the importance of web pages in web search engines. S. Brin & L. Page have used Power iteration method to find out dominant eigenvalue and eigenvector of hyperlink matrix [1]. Due to huge size of web and iterative nature of PageRank method, Power method takes several day and many resource to compute the PageRank method [3, 8, 12]. Speeding up of PageRank computation is required due to two reasons that: *one, To reduce the gapping time between a new crawl when it is finished to the crawl when it can be made accessible for searching, Second, In topic sensitive and Personalized PageRank which require computation of many PageRank vectors.*These approaches forces the need of faster method to compute PageRank [7].

A lot of work have been done in direction of speeding up the PageRank computation by using various algebraic methods. Arnoldi et al. proposed Power arnoldi-algorithm which converges faster than Power method [14]. To accelerate the convergence of Power method, some acceleration method were proposed i.e. extrapolation method [5, 9, 11], aggregation and disaggregation methods [6, 10], Adaptive method [7] and lumping of nodes [13, 15]. Arasu et al. used Gauss-Seidel and Successive over Relaxation iterative method which takes less number of iteration in the convergence to compute PageRank than Power method [3].

In this paper we have discussed about Successive Over Relaxation (SOR) method and have implemented this method to compute PageRank by Hash-map data-structure on various datasets. Further we have also analyzed the result obtained after the execution of SOR method.

This paper is organized as follows, in section 2 we discuss about basic PageRank algorithm. Section 3 discuss as the computation of PageRank method by Successive over Relaxation method. In Section 4, we do the Experimental analysis of this algorithm based on number of iteration. Finally Section 5 concludes the paper.

Basic PageRank method

The fundamental concept of PageRank method inspires from human behavior of polling. Large number of web pages referenced by other web pages by hyperlinks and established huge web graph of the internet. Web pages referenced by many other web pages will thus get high rank values and considered to be most interesting/important web pages. The PageRank method can be computed as follows: let N be the total number of web pages in the web graph and $\pi(v)$ represents the rank value of web page v, N_v be the number of web pages pointed by page v and S_v represents the set of pages pointing to page v. Brin & Page proposed the following equation to compute the PageRank value of web page v [1, 12]:

$$\pi(v) = \frac{(1-\alpha)}{n} + \sum_{u \in S_v} \frac{\pi(u)}{N_v} \qquad (1)$$

DOI: 10.141079/IJITKM.2015.910

Here α is called “damping factor” value, the transition probability of the random web user and $\pi(v)$ is the PageRank of web Page v. Equation (1) is called a linear equation that can be solved by various iterative methods: *Power method [2, 4], Gauss-Seidel method and Successive over Relaxation method [3, 12]* etc. Power method is one of oldest iterative method to compute PageRank algorithm but due to iterative nature of PageRank method it takes too much time and iteration in convergence. Gauss-Seidel method differs from power method in a way that this method implements the policy of always used the latest available value of a particular variable and takes very less time and iteration than Power method in convergence to compute the PageRank for large web dataset [3].

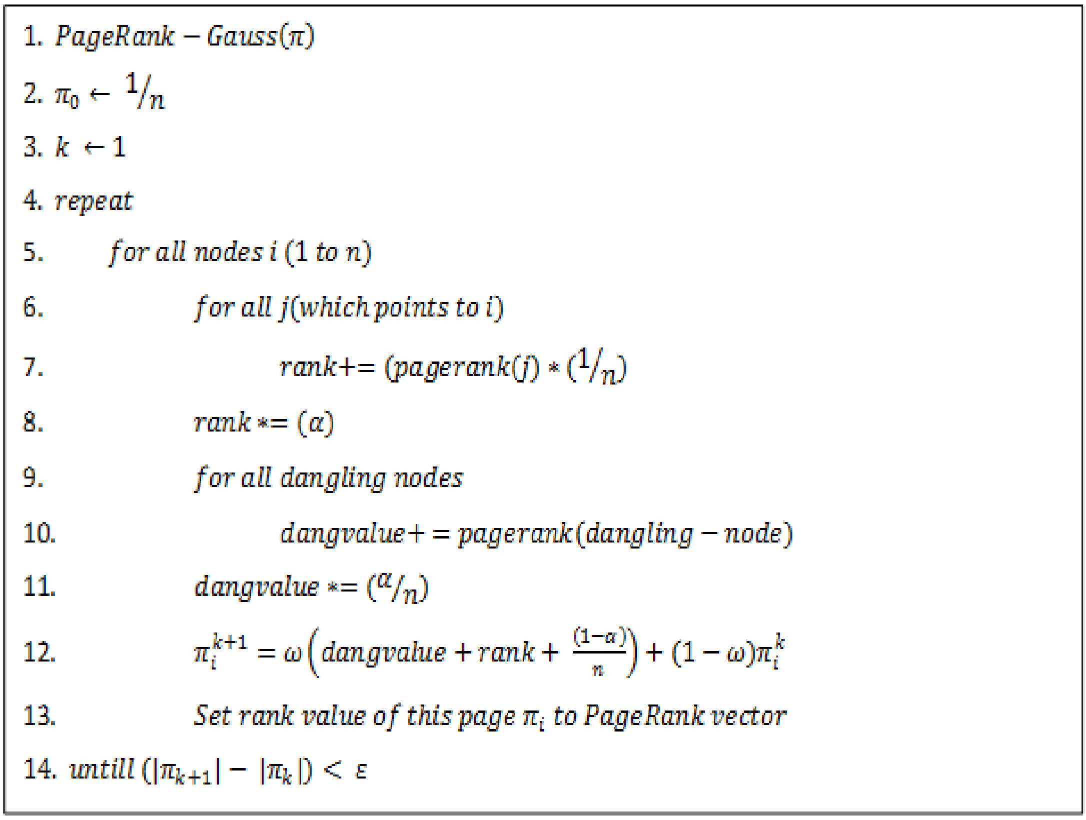
PageRank computation by SOR method

Successive over Relaxation (SOR) method is an iterative method that accelerates the convergence of Gauss-Seidel method. SOR method compute the PageRank score similar to Gauss-Seidel method only it add a *relaxation parameter* ωwhich accelerate its convergence. Arasu et al. has used the following equation to compute the PageRank algorithm [3]:

$$\pi_i^{k+1} = \omega\left\{\frac{1-\alpha}{n} + \alpha\left(\sum_{j<i} a_{ij}\pi_j^{k+1} + \sum a_{ij}\pi_j^{k}\right)\right\} + (1-\omega)\pi_i^{k} \qquad (2)$$

Here a_{ij} is the link between *web page i* to *web page j* and ω is the relaxation parameter. If value of relaxation parameter $\omega \in (0,1)$ then the iterative method is known as Successive under Relaxation and can be used to obtain the convergence when the Gauss-Seidel not converges. For $1 < \omega < 2$ this method is called Successive Over Relaxation and it is used to accelerate the convergence of Gauss-Seidel method. Also for ω=1, it simply act as Gauss-Seidel equation [3].

SOR algorithm begins with initializing *1/n* to all web pages and iterates above equation 2 till convergence reached *i.e.* for tolerance value $\epsilon = 10^{-6}$. This algorithm is based on Arasu et al. equation which we have computed on Hash-map data-structure is given below [3, 12]:



```
$PageRank - Gauss(\pi)$
$\pi_0 \leftarrow {}^{1}/_{n}$
$k \leftarrow 1$
$repeat$
    $for\ all\ nodes\ i\ (1\ to\ n)$
        $for\ all\ j(which\ points\ to\ i)$
            $rank += (pagerank(j) * ({}^{1}/_{n}))$
        $rank *= (\alpha)$
        $for\ all\ dangling\ nodes$
            $dangvalue + = pagerank(dangling - node)$
        $dangvalue *= ({}^{\alpha}/_{n})$
        $\pi_i^{k+1} = \omega\left(dangvalue + rank + \frac{(1-\alpha)}{n}\right) + (1-\omega)\pi_i^k$
        $Set\ rank\ value\ of\ this\ page\ \pi_i\ to\ PageRank\ vector$
$untill\ (|\pi_{k+1}| - |\pi_k|) < \varepsilon$
```

Figure 1: Successive over Relaxation method to compute PageRank

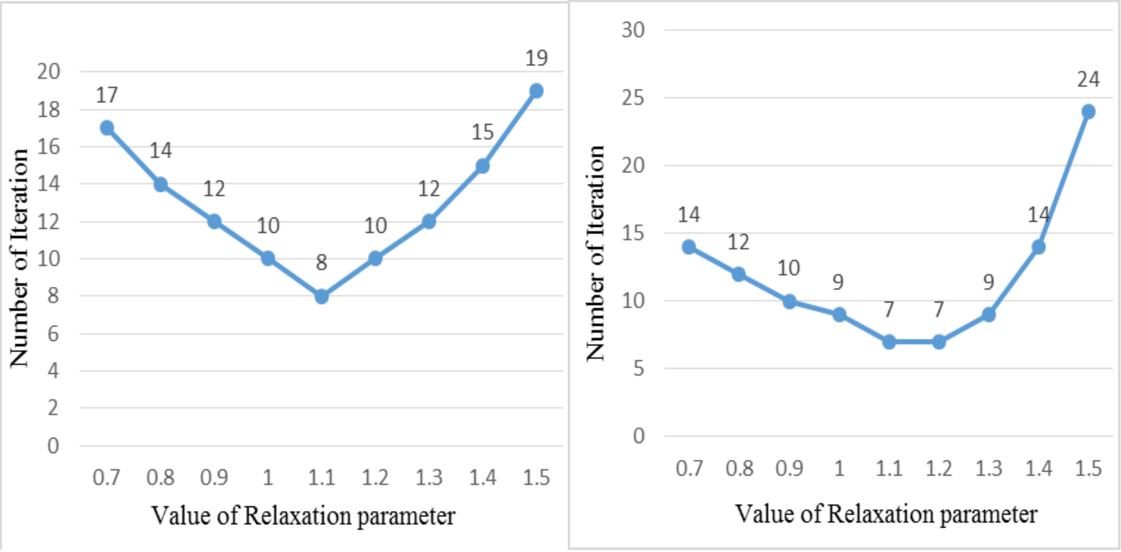
Experimental Analysis

We have implemented Successive over Relaxation method in Java language by Hash-map and ImmutableMulti-map data-structure [17]. We have done experiment on single Linux machine (Ubuntu 14.04 LTS), an Intel Core i5 CPU 3.2 GHz. The SOR method computed on the following dataset that is collected from Stanford large network dataset collection website that contains various type of network graph dataset [16].

Table 1: Description of Datasets

Dataset	Dataset Description	Number of nodes	Number of hyperlink	Dangling nodes
D1	Movie dataset	8846	26786	4996
D2	Graph of Gnutella peer-to-peer file sharing network where nodes represent host and edges corresponding to connection between hosts.	22687	54705	16466
D3	Graph of Enron email communication network where nodes and edges are corresponding to email addresses and connection between email address respectively.	36692	183831	0
D4	Graph of location based social networking service provider where users shared their locations by checking-in.	58228	214078	0

After the execution of SOR algorithm on these above dataset the results are represented by following figure 2 (a, b, c, d). From the Fig. 2 itself we can say that for $\omega < 1$ Successive Relaxation method performs as Successive under Relaxation method and takes more number of iteration than Gauss-Seidel method and for $1 < \omega < 2$ it acts as Successive over relaxation method and takes less number of iteration for any particular value of relaxation parameter ω. From Fig. 2 we can see that Successive Over Relaxation method takes less number of iteration than Gauss-Seidel method not for any particular value of relaxation parameter ω for all dataset but it's value differ for different dataset i.e. for *D1* dataset it gives an optimum number of iteration for $\omega = 1.1$; for *D2* dataset value of $\omega = 1.1\ or\ 1.2$ it gives optimum number of iteration; $\omega = 1.3\ or\ 1.4$ provide an optimum number of iteration for *D3* dataset; and for *D4* dataset it becomes $\omega = 1.4$. So we can say that for different dataset we have to find out the value of relaxation parameter ω which gives an optimum number of iteration.So we can say that for different datasets the value of relaxation parameter ω that generates an optimum number of iteration can obtained to be find out.



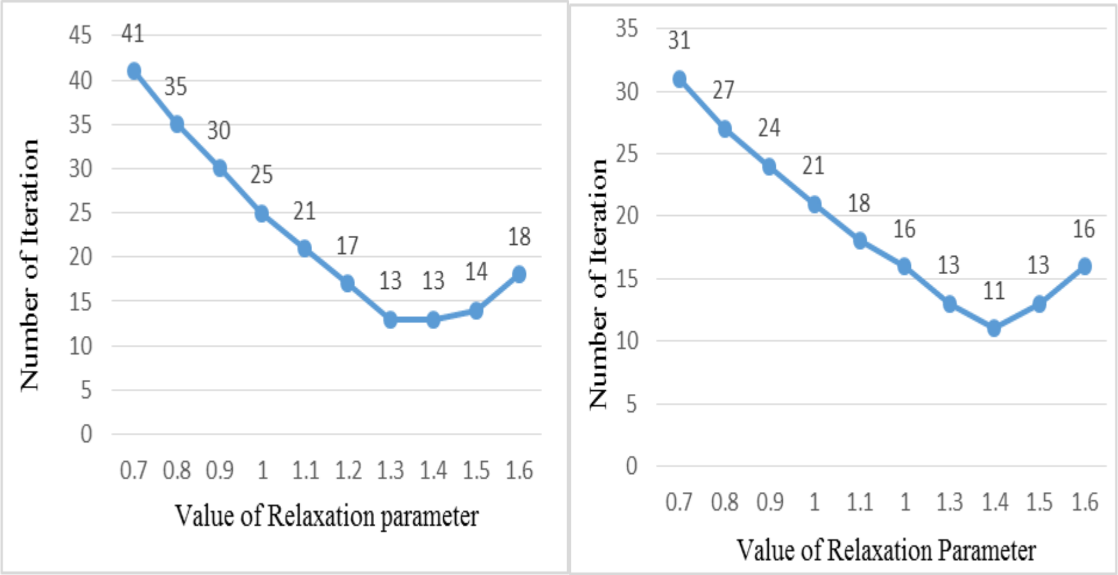


Figure 2(a) Figure 2(b)

Figure 2(c) Figure 2(d)

Figure 2: Number of Iteration taken by Successive Over Relaxation method for $0 < \omega < 2$ with damping factor value $\alpha = 0.85$ and tolerance value $\varepsilon = 10^{-6}$ for different dataset i.e. figure 2 (a), figure 2(b), figure (c) and figure 2(d) for D1, D2, D3 and D4 dataset respectively

Conclusion

Power method is the basic iterative method which used to compute the PageRank of web pages by Google search engine. Today PageRank becomes so popular and is used in various era other than web search engine e.g.: many chemist uses PageRank to emulate Chemical reaction, to determine the structure of Molecules, to destroy an ecosystem with the help of PageRank algorithm etc. Fast computation of PageRank is needed due to extensive application of PageRank. Due to iterative nature of PageRank, Power method takes many days to converge the method while Gauss-Seidel method takes less number of iteration than Power method. We have done experimental analysis of Successive over Relaxation iterative method and concludes that for large dataset it takes very less number of iteration than Gauss-Seidel method for a particular value of relaxation parameter.

References

[1] S. Brin, L. Page (1998), “The Anatomy of a Large-scale Hyper textual Web Search Engine” Proceedings of the Seventh International World Wide Web Conference, Page(s):107-117.

[2] Pavel Berkhin (2005), “A survey on PageRank computing”, Internet Mathematics 2, Vol.1, Page(s):73–120.

[3] Arasu, Arvind, et al. "PageRank computation and the structure of the web: Experiments and algorithms." Proceedings of the Eleventh International World Wide Web Conference, Poster Track. 2002.

[4] Pretto, L.: A theoretical analysis of googles PageRank. In: Laender,A.H.F., Oliveira, A.L. (eds.) SPIRE 2002. LNCS, vol. 2476, pp. 131144. Springer, Heidelberg (2002).

[5] C. Brezinski, M. Redivo-Zaglia, S. Serra-Capizzano, Extrapolation methods for PageRank computations, C. R. Acad. Sci. Paris, Ser. I 340 (2005) 393-397.

[6] I.C.F. Ipsen, S. Kirkland, Convergence analysis of a PageRank updating algorithm by Langville and Meyer, SIAM J. Matrix Anal. Appl. 27 (2006) 952-967.

[7] S.D. Kamvar, T.H. Haveliwala, G.H. Golub, Adaptive methods for the computation of PageRank, Linear Algebra Appl. 386 (2004) 51-65.

[8] Srivastava, Atul Kumar, et al. "International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS) www. iasir. net." algorithms 3.7: 14.

[9] T.H. Haveliwala, S.D. Kamvar, D. Klein, C. Manning, G.H. Golub, Computing PageRank using power extrapolation, Stanford University Technical Report, July 2003.

[10] A.N. Langville, C.D. Meyer, Updating PageRank with iterative aggregation, in: Proceedings of the Thirteenth World Wide Web Conference, ACM Press, New York, 2004, pp. 392-393.

[11] S.D. Kamvar, T.H. Haveliwala, G.H. Golub, Extrapolation methods for accelerating PageRank computations, Technique Report SCCM 03-02, Stanford University, 2003.
[12] S. Serra Capizzano, Google PageRank problem: The model and the analysis,in: Proceedings of the Dagstuhl Conference in Web Retrieval andNumerical Linear Algebra, 2007.
[13] Y. Lin, X. Shi, Y. Wei, on computing PageRank via lumping Google matrix, J. of Comput. and Appl. Math. 224 (2009) 702-708.
[14] G. Wu, Y. Wei, A power-Arnoldi algorithm for computing PageRank, Numer. Linear Algebra Appl. 14 (2007) 521-546.
[15] C.P. Lee, G.H. Golub, S.A. Zenios, A fast two-stage algorithm for computing PageRank and its extensions, Stanford University Technical Report, SCCM- 03-15, 2003.
[16] Jure Leskovec and Andrej Krevl, Stanford Large Network Dataset Collection, http://snap.stanford.edu/data, june-2014.
[17] http://code.google.com/p/guava-libraries/

DOI: 10.141079/IJITKM.2015.910