Quadratic-Extrapolated Gauss-Seidel Method to Accelerate PageRank Computation

Atul Kumar Srivastava*, Rakhi Garg** and P. K. Mishra***
*, ***Department of Computer Science, Institute of Science
**Department of Computer Science, Mahila Maha Vidyalaya
Email: atulbhuphd@gmail.com

Abstract: PageRank is a basic metric used to measure the importance of web pages. It is an iterative method, and computed by Power method. Power method converges too slowly for the true structure of web. This issues encourages to find out some other iterative methods to accelerate PageRank computation. In this paper, we present a novel method that is combination of Quadratic Extrapolation and Gauss-Seidel called QE-GSeidel method which accelerate the convergence of PageRank method. From numerical experiment we show that QE-GSeidel method takes almost 60-65% less number of iteration and 25-30% less CPU time than Power, and takes 50-55% less number of iteration and 15-20% less CPU time than Gauss-Seidel in PageRank computation.

Keywords: PageRank, Power method, Gauss-Seidel method, Quadratic Extrapolation method.

Introduction

The PageRank algorithm is used to compute the "importance" or ranking of web pages. Now, it has become a basic method for web search engine to compute the rank of pages [1, 2]. The basic idea of the PageRank algorithm is to compute the principal eigenvector of hyperlink matrix of web graph. Due to exponential growth of web, PageRank computation becomes very crucial task and it takes several days in computation. There is requirement to accelerate the computation of PageRank so that search engine respond to fast query result. Power method is one of the basic numerical method that is used to compute PageRank algorithm [2]. However, it is well known that convergence rate of the power method is very slow when the damping factor value close to one [3, 4, 5]. Therefore it becomes necessary to accelerate the convergence of PageRank computation by developing some efficient method.

Recently there are many researchers try to explore some other efficient numerical method to accelerate the convergence of the PageRank computation. Kamvar et al, elaborate adaptive method to accelerate the convergence of PageRank computation [6]. In this method PageRank of those pages are not recomputed at every iteration that are already converges in previous iterations. This method accelerate the convergence of PageRank method but the pages that converge in earlier iteration gives inappropriate result. The Power-Arnoldi PageRank algorithm combines the power method with the thick restarted Arnoldi extrapolation algorithm to speedup PageRank computation [7]. The Arnoldi type algorithm is based on restarted krylov subspace method. It combines the Arnoldi process and singular value decomposition of larger eigenvalues [7, 8]. Among these strategies, extrapolation acceleration methods have been implemented constantly. In 2006, Sependar et al, proposed an extrapolation techniques that off estimate the eigenvalues to accelerate the convergence of power method to compute PageRank algorithm [9]. However, this method does not perform efficiently when extrapolation applied very frequently, and for damping factor close to one. Motivated by these research, we produce a new extrapolation method called Quadratic Gauss-Seidel *i.e.* the combination of Quadratic extrapolation [9] and Gauss-Seidel method to speed up the convergence of Gauss-Seidel method. Furthermore, this combined method performs very efficiently in PageRank computation. In this paper, we combined Quadratic extrapolation with Gauss-seidel to compute PageRank problem. This algorithm is called as quadratic-extrapolated Gauss-Seidel that accelerates the convergence of PageRank method. The rest of the paper is organized as follows: In Section 2, discuss PageRank computation. In section 3, we discuss quadratic-extrapolation Gauss-seidel method based on hyperlink matrix then implement it to compute the PageRank. The numerical experiment is provided in section 4 that observes the numerical behavior of proposed algorithm. Finally, in section 5 some concluding remarks and future works are presented.

Preliminaries: PageRank Method

In this section we discuss the basic definition of PageRank and how it is computed by using numerical methods. The basic definition of PageRank state that "A web page is important or get higher rank if it is pointed by other important web pages". Let the hyperlink matrix $H \in R^{n \times n}$ of the web graph defined as following equation (1) [2]:

$$H\left(h_{ij}\right) = \begin{cases} {}^{1}/_{Outdeg(i)} & , if\ page\ i\ points\ to\ j \\ 0 & , otherwise \end{cases} \qquad (1)$$

In equation (1), $Outdeg(i)$ denotes number of out-going links of page i. If a web page i does not contain any out-going link, then corresponding entry *i.e.* i^{th} row of H will be zero and page i is called as dangling web pages. Dangling web pages create problems in the PageRank computation. Due to these web pages we can't compute unique PageRank vector for web graph. To solve this issues, simply convert the hyperlink matrix H to H' by following equation (2) [3, 4]:

$$H' = H + d(z^T) \qquad (2)$$

In equation (2), $z \geq 0, z_i \in R^{n\times n}$ with $\| z \|_1 = 1$ and $d = d_i \in R^{n\times n}$ with

$$d_i = \begin{cases} 1, & if\ outdeg(i) = 0 \\ 0, & otherwise \end{cases} \qquad (3)$$

Now H' becomes a row stochastic matrix *i.e.* $H'e = e$, where e is an identity vector of $n \times 1$ *i.e.* $e = (1, 1, 1, \ldots, 1)^T$. To ensure the irreducibility and aperiodicity of H', a damping factor α and a personalization vector p are introduced to establish the final matrix M as in following equations.

$$M = (\alpha H' + (1-\alpha).e.p^T)^T \qquad (4)$$

$$M = \left(\alpha H'^T + (1-\alpha).p.e^T\right) \qquad (5)$$

Now matrix M becomes column stochastic and irreducible. By *Perrorn-Frobenus [2, 3]*, matrix M gives maximum eigenvalue equal to one, with corresponding eigenvector is non-negative and unique. After normalization of this vector, it is called as PageRank vector π, satisfying the following equation (6) [4, 5]:

$$M.\pi = \pi, \qquad \| \pi \|_1 = 1 \qquad (6)$$

Equation (6) is an iterative method and it is implemented by basically Power method in PageRank computation [2, 3]. Many researchers have used several other algebraic method like Gauss-Seidel, Monte Carlo, SOR (Successive Over Relaxation) method to compute equation (6) [5, 10, 11]. In this paper we compute PageRank by using Gauss-Seidel method (Algorithm 1) then we will apply quadratic extrapolation on it to accelerate the convergence of Algorithm (1). According to Gauss-Seidel algorithm for a starting PageRank vector π^0, the iterative Sequence vector *i.e.* π^l converges to a unique PageRank vector of matrix M .

Algorithm 1: PageRank computation using Gauss-Seidel [11]

Input :

π^0 = Initial PageRank column vector zeroth iteration
n = Number of web pages in web graph
α = Damping factor
δ = Tolerance value
O_j = [1/Out-degree of j^{th} web page]
l = Number of iteration
m = Dangling node

Output: π = Final PageRank column vector

```
$\pi^0 = {}^{1}/_{n}$ // initialize PageRank vector
$l$ = 0
while ($tole \leq \delta$)
        $l$++
        $\pi_i^l = \alpha\left[\sum_{j<i} \pi_j^l O_j + \sum_{j>i} \pi_i^{l-1} O_j\right] + \frac{1-\alpha}{n} + \frac{\alpha}{n}\left[\sum_{\forall m<i} \pi_m^l O_m + \sum_{\forall m>i} \pi_m^{l-1} O_m\right]$
        $tole = norm\, |\pi^l - \pi^{l-1}, 1|$
        $if$ converged $then\ \pi^l$
end
return $\pi^l$
```

Due to slow convergence of Power method, it is necessary to find out some fast numerical methods to compute PageRank. Hence it is important to design a simple and efficient algebraic numerical method that can accelerate convergence of PageRank algorithm. In the following section we applied Quadratic Extrapolation technique to Gauss- Seidel method to compute the PageRank value of matrix *M*, and we show by experimentally that it accelerate the convergence of Gauss-Seidel method to compute PageRank of web pages.

Quadratic Extrapolated Gauss-Seidel PageRank Method

We combined Quadratic extrapolation with Gauss-Seidel method to accelerate the convergence of PageRank computation [9]. We assume that matrix M has three eigenvectors$\vec{m}_1$, $\vec{m}_2$ and $\vec{m}_3$, although the matrix has more than 3 eigenvectors, and the iteration π^{l-3} can be defined as the combination of these three by following equation (7):

$$\pi^{l-3} = c_1\vec{m}_1 + c_2\vec{m}_2 + c_3\vec{m}_2 \tag{7}$$

In equation (7), $c_1 = 1, c_2, and\ c_3$ are constant. The successive iterations can be solved by using equation (7) as follows:

$$\pi^{l-2} = M.\pi^{l-3} \tag{8}$$

$$\pi^{l-1} = M.\pi^{l-2} \tag{9}$$

$$\pi^{l} = M.\pi^{l-1} \tag{10}$$

We compute the value of $\vec{m}_1, \vec{m}_2, and\ \vec{m}_3$ by using Cayley-Hamilton theorem. Algorithm (2) shows that how we applied Quadratic Extrapolation in Gauss-Seidel method to accelerate the convergence of PageRank algorithm.

Algorithm 2: PageRank computation using Quadratic-Extrapolated Gauss-Seidel method

Compute PageRank value of every web page of n*n hyperlink matrix *H* with initial PageRank vector π^0, given damping factor value α, and tolerance norm δ. We applied quadratic extrapolation at every *t* iteration.

Input :
π^0 = Initial PageRank column vector zeroth iteration
n = Number of web pages in web graph
α = Damping factor
δ = Tolerance value
O_j = [1/Out-degree of j^{th} web page]
l = Number of iteration
m = Dangling node

Output: π = Quadratic Gauss-Seidel method

```
$\pi^0 = 1/n$ // initialize PageRank vector
$l = 0$
while ($tole \le \delta$)
    $l$++
    $copy\ (curr\_\pi, \pi)$
    $curr\_\pi_i^l = \alpha\left[\sum_{j<i}\pi_i^l O_j + \sum_{j>i}\pi_i^{l-1} O_j\right] + \frac{1-\alpha}{n} + \frac{\alpha}{n}\,[\sum_{\forall m<i}\pi_m^l O_m + \sum_{\forall m>i}\pi_m^{l-1} O_m]$
    $tole = norm\,|curr\_\pi^l - curr\_\pi^{l-1}, 1|$
    $while\ (l\%t == 0)$
        // begin Quadratic extrapolation procedures.
        Copy ($next\_\pi$, $curr\_\pi$)
        $next\_\pi_i^l = \alpha\left[\sum_{j<i} next\_\pi_i^l O_j + \sum_{j>i} next - \pi_i^{l-1} O_j\right] + \frac{1-\alpha}{n} + \frac{\alpha}{n}\,[\sum_{\forall m<i} next\_\pi_m^l O_m + \sum_{\forall m>i} next\_\pi_m^{l-1} O_m]$
        Copy ($nextnext\_\pi$, next_$\pi$)

        $nextnext\_\pi_i^l = \alpha\left[\sum_{j<i} nextnext\_\pi_i^l O_j + \sum_{j>i} nextnext\_\pi_i^{l-1} O_j\right] + \frac{1-\alpha}{n} + \frac{\alpha}{n}\,[\sum_{\forall m<i} nextnext\_\pi_m^l O_m + \sum_{\forall m>i} nextnext\_\pi_m^{l-1} O_m]$
        //Calculate differences
        $curr\_D = (curr\_\pi) - \pi$
        $next\_D = (next\_\pi) - \pi$
        $nextnext\_D = (nextnext\_\pi) - \pi$
        $\Upsilon = [next\_D, curr\_D]$
        $\gamma_3 = 1$
        $\binom{\gamma_1}{\gamma_2} = -\Upsilon^{+} nextnext\_D$ //Compute pseudo inverse using $QR$ Factorization
```

```
            $\beta_0 = \gamma_1 + \gamma_2 + \gamma_3$
            $\beta_1 = \gamma_2 + \gamma_3$
            $\beta_3 = \gamma_3$
            $final\_\pi = \beta_0 curr\_\pi + \beta_1 next\_\pi + \beta_2\_nextnext\pi$
        copy ($\pi$, $final\_\pi$)
    end
    return $final\_\pi$
```

From algorithm (2) it is clear that this method takes *O (n)*number of iteration.* It has minimal overhead over Gauss-Seidel method at every "a" iteration but in our experimental result it shows that this minimal overhead reduced 60% number of iteration as well as 30% time required to converge PageRank algorithm than Power method while 40% number of iteration and 26% time than Gauss-Seidel method.

Experimental Result

In this section, we have done numerical experiment that observe the performance of Quadratic extrapolated Gauss-seidel method in the PageRank Computation. All experiment have done in JAVA (JDK1.7) on an Intel$^{(R)}$ Core^{TM} i5 CPU with 6 GB RAM. In all experimental result we denote Quadratic extrapolated Gauss-Seidel, Gauss-Seidel as QE-GSeidel, GSeidel respectively. We have performed the experiment on following dataset [12]:

Table 1: Datasets with their attributes

Dataset	No. of Web pages	No. of Hyperlinks	No. of Dangling Web pages
Dataset (D1)	133279	396160	114507
Dataset (D2)	325729	1497134	187788

In all the experiment we have taken value of damping factor $\alpha = \{0.6, 0.7, 0.8, 0.85, 0.9, 0.95\}$. Figure (1) and (2) shows the number of iteration and CPU time taken to converge the PageRank algorithm using QE-GSeidel, GSeidel, and Power method. QE-GSeidel method reduces the number of iteration needed to converge the PageRank computation for tolerance value 10^{-9} and 10^{-8} by 50% and 40% approximate respectively than GSeidel method. Figure (1) shows the QE-GSeidel takes almost 18% and 26% less time than GSeidel method and Power method to reach the convergence value of 10^{-9}, and for tolerance value 10^{-8}, it reduces almost 14% and 22% time than GSeidel and Power method. From figure 2(a), to reach a tolerance vale 10^{-9}, QE-GSeidel saved 18% time over GSeidel method and 30% time over Power method. An important analysis of QE-GSeidel method is that it is necessary that we apply extrapolation very frequently. In these experiment we applied Quadratic extrapolation at every 10th iteration. So form figure 1 and 2, it is clear that although QE-GSeidel have some overhead in the computation with Gauss-Seidel method but it converges faster than Power method and Gauss-Seidel method in terms of CPU time as well as Number of iteration.

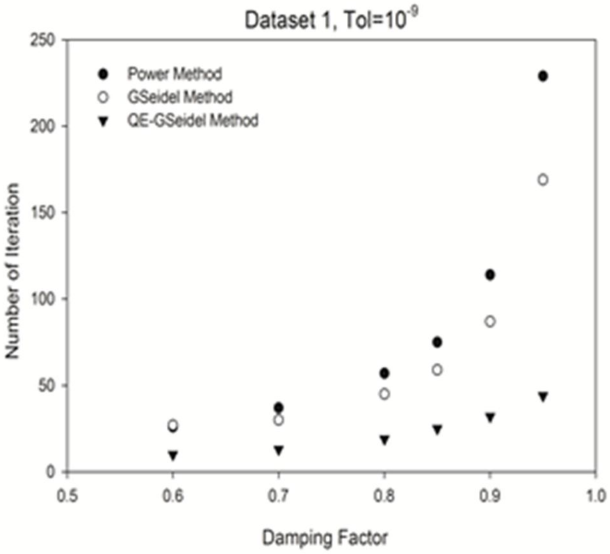

Fig. 1(a)

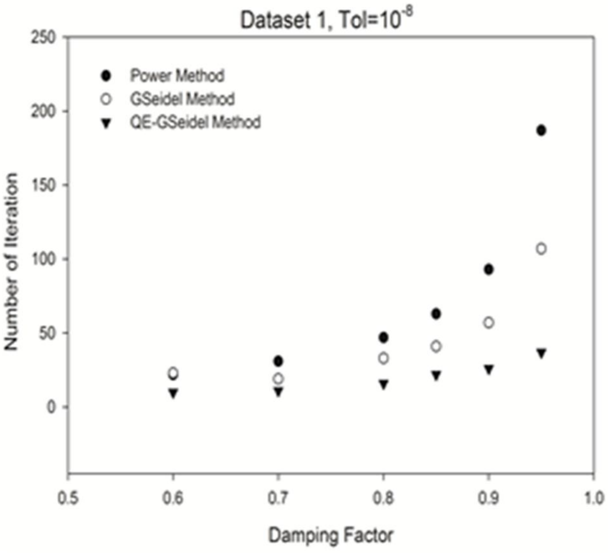

Fig. 1 (b)

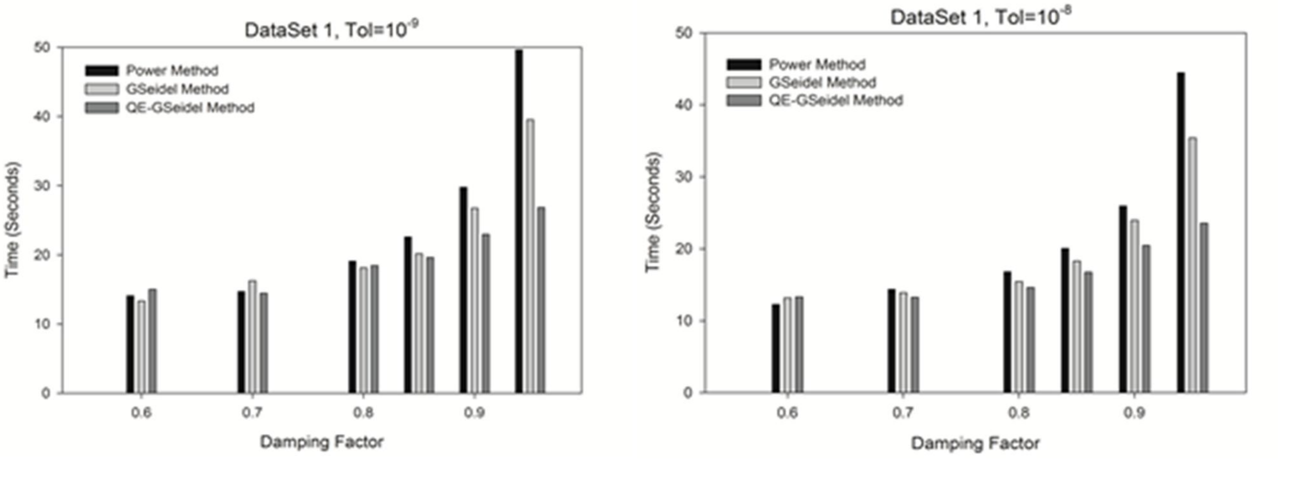

Fig. 1(c)

Fig. 1 (d)

Fig. 1: Graph shows the number of iteration and time taken in the PageRank computation for various tolerance value of D1 dataset

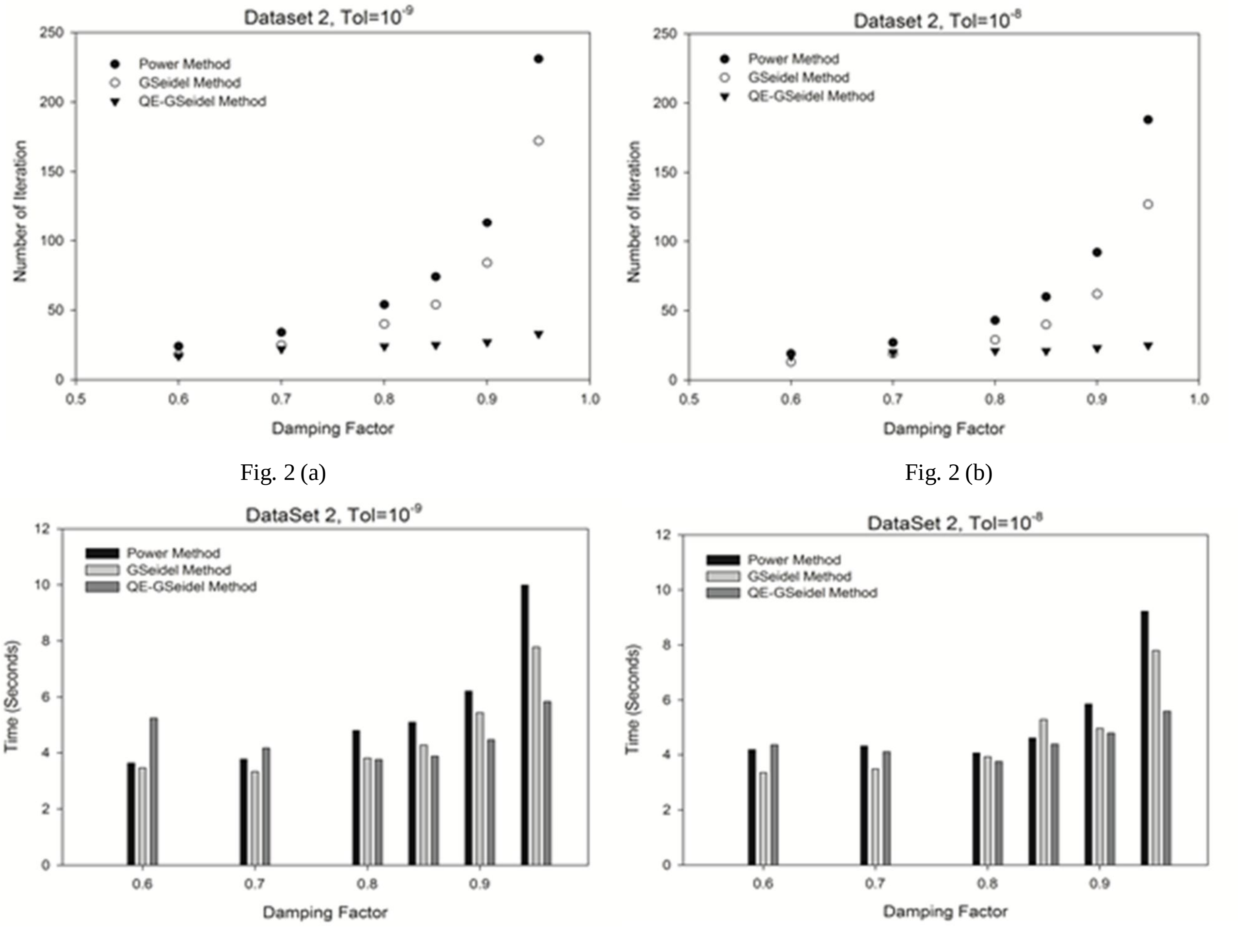

Fig. 2 (a)

Fig. 2 (b)

Fig. 2 (c)

Fig. 2 (d)

Fig. 2: Graph shows the number of iteration and time taken in the PageRank computation for various tolerance value of D2 dataset

Conclusion

Web Search engines has become these days an integral part of information access today, that possess many interesting issues for the developers to provide an effective and efficient method to rank the web pages. PageRank is one of the famous web page ranking algorithm that computes rank based on authoritativeness of web pages by using hyperlink structure of web. Currently, exponential growth in web crawl repository, increase in crawling frequency, and Personalized PageRank required to speeding up the computation of PageRank. Quadratic extrapolation is an extrapolation technique of linear equation system that integrated into Gauss-Seidel method. In this method we applied Quadratic Extrapolation periodically in Gauss-Seidel to accelerate the convergence of PageRank. From experiment it shows that this method takes % less number of iteration and % less CPU time than basic PageRank Power method.

References

[1] P. Berkhin, “A survey on pagerank computing,” Internet Math., vol. 2, no. 1, pp. 73–120, 2005.
[2] L. Page, S. Brin, R. Motwani, and T. Winograd, “The PageRank citation ranking: Bringing order to the web.” 1999.
[3] A. N. Langville and C. D. Meyer, Google’s PageRank and beyond: The science of search engine rankings. Princeton University Press, 2011.
[4] 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.
[5] A. N. Langville and C. D. Meyer, “Deeper inside pagerank,” Internet Math., vol. 1, no. 3, pp. 335–380, 2004.
[6] S. Kamvar, T. Haveliwala, and G. Golub, “Adaptive methods for the computation of PageRank,” Linear Algebra Appl., vol. 386, pp. 51–65, 2004.
[7] G. Wu and Y. Wei, “An Arnoldi-extrapolation algorithm for computing PageRank,” J. Comput. Appl. Math., vol. 234, no. 11, pp. 3196–3212, 2010.
[8] Wu, G., & Wei, Y. (2007). A Power–Arnoldi algorithm for computing PageRank. Numerical Linear Algebra with Applications, 14(7), 521-546.
[9] Kamvar, S. D., Haveliwala, T. H., Manning, C. D., & Golub, G. H. (2003, May). Extrapolation methods for accelerating PageRank computations. In Proceedings of the 12th international conference on World Wide Web (pp. 261-270). ACM.
[10] M. Brinkmeier, “PageRank revisited,” ACM Trans. Internet Technol., vol. 6, no. 3, pp. 282–301, 2006.
[11] A. Arasu, J. Novak, A. Tomkins, and J. Tomlin, “PageRank computation and the structure of the web: Experiments and algorithms,” in Proceedings of the Eleventh International World Wide Web Conference, Poster Track, 2002, pp. 107–117.
[12] J. Leskovec and A. Krevl, “{SNAP Datasets}: {Stanford} Large Network Dataset Collection.” Jun-2014.