Computer Engineering and Applications Vol. 1, No. 2, December 2012

Using Power-Law Degree Distribution to Accelerate PageRank

Zhaoyan Jin[1], Quanyuan Wu[2]
School of Computer
National University of Defense Technology
Changsha China 410073
Email:[1]jinzhaoyan@163.com
[2]wqy.nudt@gmail.com

ABSTRAKSI

Sebuah vektor pagerank jaringan sangat penting. Hal ini dapat menggambarkan pentingnya suatu halaman web pada jaringan web global, atau seseorang pada media sosial. Namun dengan perkembangan dari jaringan web global dan media sosial. Hal ini membutuhkan lebih banyak waktu untuk menghitung vector pagerank dari sebuah jaringan. Dalam banyak aplikasi dunia nyata, distribusi derajat dan PageRank dari jaringan yang kompleks sesuai dengan distribusi Power-Law. Makalah ini memanfaatkan distribusi derajat jaringan untuk menginisialisasi vektor PageRank, dan menyajikan algoritma percepatan distribusi derajat power-law dari perhitungan pagerank. Percobaan pada empat himpunan data dunia nyata menunjukkan algoritma yang diusulkan berkonvergen lebih cepat dibandingkan dengan algoritma asalnya.

Kata kunci: pagerank, media social, power-law, distribusi derajat

ABSTRACT

The PageRank vector of a network is very important, for it can reflect the importance of a Webpage in the World Wide Web, or of a people in a social network. However, with the growth of the World Wide Web and social networks, it needs more and more time to compute the PageRank vector of a network. In many real-world applications, the degree and PageRank distributionsof these complex networks conform to the Power-Law distribution. This paper utilizes the degree distribution of a network to initialize its PageRank vector, and presents a Power-Law degree distribution accelerating algorithm of PageRank computation. Experiments on four real-world datasets show that the proposed algorithm converges more quickly than the original PageRank algorithm.

Keywords: PageRank, Social Network, Power-Law, Degree Distribution

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

1. INRODUCTION

One of the fundamental problems in information retrieval is the ranking problem: given a query, how to arrangethe documents which satisfy the query such that the most relevant ones rank first. Before the World Wide Web, the information retrieval community utilizes similarity measures to rank documents. When a query of several keywords is issued, the documents which are most similar to the query are returned.

In addition to structured keywords, Web pages also contain hyperlinks among each other, which can be thought of peer endorsements among these Web pages. With these hyperlinks, Web pages can be considered as a sparse graph. Thus, the link-based ranking algorithms, such as PageRank[1], HITS[2], SALSA[3], etc., which take peer endorsements into account, come up. Among these link-based ranking algorithms, PageRank, proposed by Google, is the most successful.

The PageRank vector$\pi$of a graph is the stationary distribution of a random walk that at each step, jumps to a random node $r$with the probability ε, and follows a random outgoing edge from the current node with the probability 1-ε. Given a weighted directed graph G=(V,E) with n nodes and m edges, the weight on an edge $(u,v)\in E$ is denoted with $a_{u,v}$.The Transition ProbabilityMatrix $P=\{p_{i,j}\}$ of G is defined as follows:

$$p_{i,j}=\begin{cases}\dfrac{a_{i,j}}{\sum_{(i,k)\in E}a_{i,k}}, & a_{i,j}\neq 0\\ 0, & a_{i,j}=0\end{cases} \tag{1}$$

The PageRank vectorπ of a graph satisfies:

$$\pi=(1-\varepsilon)P^T\pi+\varepsilon\cdot r \tag{2}$$

whereP^T is the transpose of transition probability matrix$P$, and $\boldsymbol{r}$ is a vector each of which$r(i)$ is a probabilitywith which the random walk jumps to the node i.

There are a number of numerical methodsfor computing the PageRankvector. However, in spite of its low efficiency,the power iterationmethod[1]stands outfor its stable and reliable performances. It starts with initializing $\pi^{(0)}(v)=r(v)$ (where v is arbitrary node in the graph), and then performsformula 2 repeatedly until it converges, i.e., the difference between two consecutive iteration is under a certain constantβ. To remedy the slow convergenceof the power iteration method, several acceleration techniques have been proposed, whichinclude extrapolation[4], aggregation[5], lumping[6], and adaptive methods[7]. Moreover, the Arnoldi-type method is introduced by Gene et al.[8], and the Jordan canonical form of the Google matrix isinvestigated by Wu[9].

The Power-Law distribution is an important characteristic about distribution of nodes' degrees in complex networks, such as the World Wide Web and social networks. Meanwhile, the

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

PageRank vectors of these networks also conform to the Power-Law distribution.The Power-Law distribution can be described as follows:

$$f(x) \propto e^{\lambda} \tag{3}$$

Where x is degree or PageRank, λ is exponent or scaling parameter, and $f(x)$ means the number of nodes having that degree or PageRank respectively.This paper utilizes the degree distribution of a network to initialize its PageRank vector, and presents a Power-Law degree distribution method for accelerating the PageRank computation.

2. POWER-LAW DEGREE DISTRIBUTION ACCELERATING METHOD OF PAGERANK COMPUTATION

In the physical world, the Power-Law distribution is animportant characteristic in complex networks. To illustrate this idea, this paper presents four real-world social networks, Dianping[11], Wikipedia-Film[10], Epinions[12]and Gowalla[13].Some statistics of these datasets are in Table 1.

Table 1.Statistics of Datasets

Name	# nodes	# edges	λ(degree)	λ(PageRank)
Wikipedia-Film	27312	142427	1.989	1.625
Epinions	75879	508837	1.593	1.678
Gowalla	196591	1900654	1.744	1.759
Dianping	204074	926720	1.822	2.1

The degree or PageRank distribution of a graph clarifiesthat the number of nodes changes with the node's degree or PageRank score respectively.The PageRank score of $v$denoted with$\pi(v)$ is a decimal fractionbetween 0 and 1. The degree distributions and PageRank distributions of these four graphs are in Figure 1. This paper first finds the minimum score $\pi_{min}$and the maximum PageRank score π_{max}in each graph, and then changes each PageRank score into an integer according to the following formula:

$$\pi(v)' = 1000 \cdot \frac{\pi(v) - \pi_{min}}{\pi_{max} - \pi_{min}} \tag{4}$$

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

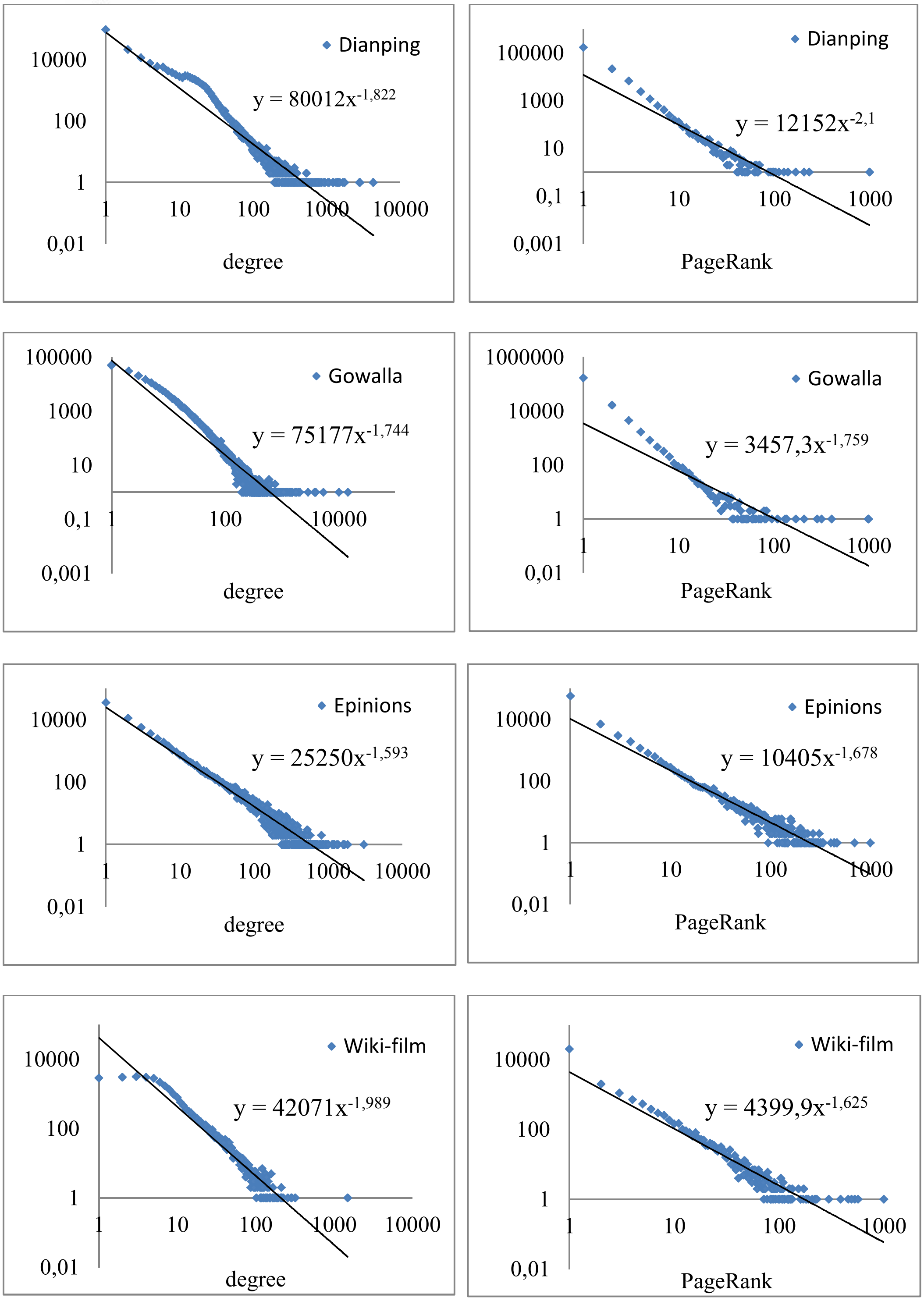

Figure 1. The Power-Law Distribution

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

As can be seen from Figure 1 that, the degree distributions and the PageRank distributions of these graphsall conform to the Power-Law distribution.Based on this fact, this paper proposes the Power-Law degree distribution method to accelerate the convergence of the power iteration PageRank computation, details of the proposed algorithm is in Algorithm 1.

Algorithm 1 PowerLawDegreePageRank(G)

Input: a graph G (V, E), where $|V|$=n and $|E|$=m;
Output: a list of(i,π(i)), where a node is denoted with$i$($i \in V$ and $1 \le i \le n$) and its PageRank scoreis denoted with π(i);

for i in 1 to n
let *neighbor*=number of neighbors of i;
  let d_i= *neighbor/2m*;
end for
letπ=**0**, π'= ***d***;
while |π'-π|>β**do**
  let π=π';
for$i$ in 1 to n
$\pi'(i) = (1\text{-}\varepsilon)\sum_{(k,i)\in E} \frac{\pi(k)}{|outlink(k)|} + \frac{\varepsilon}{\text{n}}$;
end for
end while

There are three steps in this algorithm. Firstly, compute the degree distribution vector ***d***, i.e., count the number of neighborsd_i (include in-links and out-links) for each node $i\,(1 \le i \le n)$; secondly, initialize the PageRank vector with the degree distribution vector ***d***,and then normalize it according to $\pi^{(0)}(i) = \frac{d_i}{2m}$; thirdly, compute the PageRank vector repeatedly with formula 2 until it converges.

3. EXPERIMENTS

This section describes the results ofthe experiments that we have done to validate the efficiency of the proposed method. The experiments are done on a personal computer, and the algorithmsare implemented on JDK 1.6 and Jung 2.0.1[1].The datasets for these experiments are Dianping, which hasbeen crawled from the Dianping[2] websiteourselves, and three public datasets,Wikipedia-Film, Epinions and Gowalla.Details of these datasets are in section 2.

---

[1]http://jung.sourceforge.net/
[2] www.dianping.com

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

To validate the efficiency of the proposed method, this paper compares the proposed algorithm with theoriginal PageRank power iteration method. When a query of information retrieval is issued by a user, the user only cares about the top k results that returned. This paper chooses the 1000 PageRank iterations as the stationary distribution, and validates the similarity between each result with the stationary distribution. The top100 similarity is the number of elements which are the intersection between each result with the stationary distribution. In these experiments, the parameters are β=0.01 and ε=0.15, and details of the results are in Figure 2.

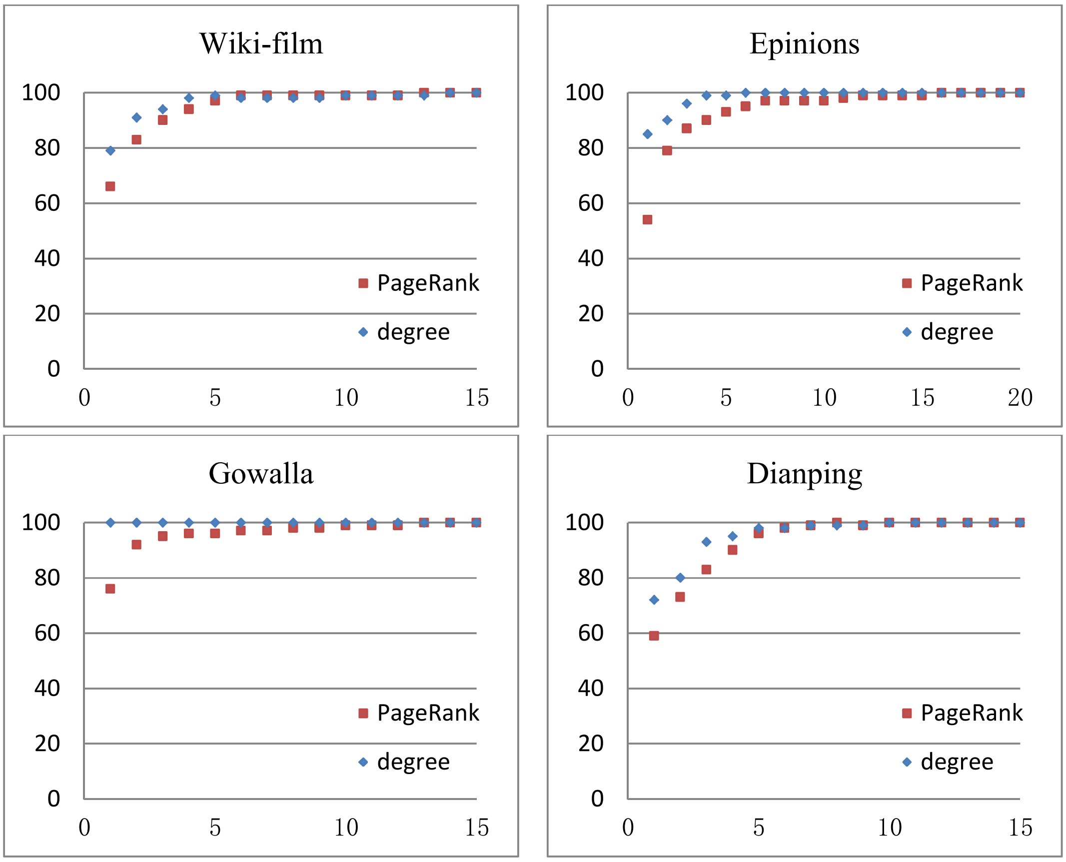

Figure 2. Comparison of Efficiency, where the x axis is the number of iteration, and the y axis is the top100 similarity with the stationary distribution.

As can be seen from Figure 2 that, the proposed Power-Law degree distribution acceleratingmethod of PageRank computation performs better than the original PageRank computation.In the Gowalla dataset, the proposed algorithm converges only at a single iteration, and in the other three datasets, the proposed algorithm converges more quickly than the original PageRank computation, too. In addition, as there may be many stationary distributions in a sparse graph, i.e., many stationary solutions for formula 2, the proposed algorithm and the original PageRank algorithm converge to different PageRank vectors in our datasets.

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

4. CONCLUSION

In many real-world complex networks, such as the World Wide Web and social networks, the degree and PageRank distributions both conform to the Power-Law distribution. This paper utilizes the degree distribution of a network to initialize its PageRank vector, and presents a Power-Law degree distribution acceleratingalgorithm of PageRank computation. Experiments on four real-world datasets show that the proposed algorithm converges more quickly than the original PageRank algorithm.In addition, the proposed algorithm can also work together with other methods to accelerate the PageRank computation further. However, the proposed algorithm performs better only in the networks that conform to the Power-Law distribution.

ACKNOWLEDGEMENTS

This work was supported in part by the National Significant Science and Technology Special Project of China (Nos. 2011ZX03002-004-01 and 2009ZX01043-002-004) and the National Natural Science Foundation of China (No. 90818028.)

REFERENCES

[1] Page L, Brin S, Motwani R, et al.The PageRank citation ranking: Bringing order to the web. Technical Report, SIDL-WP-1999-0120, Stanford InfoLab, 1999.
[2] Kleinberg J M. Authoritative sources in a hyperlinked environment. Journal of the ACM (JACM). Vol. 46, No. 5, pp. 604-632, 1999.
[3] Lempel L R, Moran S. The stochastic approach for link-structure analysis (SALSA) and the TKC effect6[J]. Computer Networks.Vol. 33, pp. 387-401, 2000.
[4] Brezinski C, Redivo-Zaglia M, Serra-Capizzano S. Extrapolation methods for PageRank computations. Comptes Rendus Mathematique. Vol. 340, No. 5, pp. 393-397, 2005.
[5] Ipsen I C F, Kirkland S. Convergence analysis of a PageRank updating algorithm by Langville and Meyer[J]. SIAM journal on matrix analysis and applications.Vol. 27, No. 4, pp. 952-967, 2006.
[6] Lin Y, Shi X, Wei Y. On computing PageRank via lumping the Google matrix. Journal of Computational and Applied Mathematics. Vol. 224, No. 2, pp. 702-708, 2009.
[7] Kamvar S, Haveliwala T, Golub G. Adaptive methods for the computation of PageRank. Linear Algebra and its Applications. Vol. 386, pp. 51-65, 2004.
[8] Golub G H, Greif C. Arnoldi-type algorithms for computing stationary distribution vectors, with application to PageRank. Technical Technical Report SCCM-04-15, Stanford University Technical Report, 2004.
[9] Wu G, Wei Y. Comments on" Jordan Canonical Form of the Google Matrix". SIAM Journal on Matrix Analysis and Applications. Vol. 30, pp. 364, 2008.

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

[10] Tang J, Sun J, Wang C, et al. Social influence analysis in large-scale networks.Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. pp.807-816, 2009.
[11] Jin Z, Shi D, Yan H, et al. LBSNRank: Personalized PageRank on Location-based Social Networks[C]. In: 4th International Workshop on Location-Based Social Networks.ACM, 2012.
[12] Richardson M, Agrawal R, Domingos P. Trust management for the semantic web. The Semantic Web-ISWC. pp. 351-368, 2003.
[13] Cho E, Myers S A, Leskovec J. Friendship and mobility: User movement in location-based social networks.Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining.pp. 1082-1090, 2011.

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)