ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 7, July 2014

DOI: 10.15662/ijareeie.2014.0307005

Advances of Novel PageRank Algorithm and Its Application

Haoxiang Wang

Meng Student, Dept. of ECE, Cornell University, Ithaca, NY, USA

ABSTRACT: Determining the influence of academic research is an important work in academe. In this paper, the co-author and the citation networks are built to determine the influence of a researcher and a research paper in the way of networks separately, and the further application is discussed. Firstly, the co-author network is built based on Erdos1.htm to determine the influence of Erdos' co-authors. Secondly, based on the citations among the papers in the database, the citation network comes into being. Thirdly, the method is implemented and basic discussion is conducted. As the final, the analysis of strengths and weaknesses is conducted.

KEYWORDS: PageRank; Text Analysis; Web Crawler; Damping Factor; Academic Impact

I.INTRODUCTION

In academic research, one researcher's influence and impact can be refected from his or her research papers. In addition, the citation and co-author are important decisive factors. The measurement of researcher's influence and impact is important both to the researcher himself and the academe. It could be found that there exists many measurement of academic impact. The impact factor(IF) of an academic journal is a measurement reflecting the average number of citations to recent articles published in the journal, which was devised by Eugene Garfield, the founder of the Institute for Scientific Information in 1975[1]. In addition, there are some other measurement such as Science Citation Index(SC1) [2], H-factor[3], etc. Moreover, the citation and co-author data can construct a huge network of academicresearch. Since network science is a hot topic nowadays, which is indeed quite useful in many fields and is convenient to detect the interactions and the structure, therefore, the citation or co-author networks can built for better measurement. With the development of data pool and size, machine learning based algorithms and tools are being used in many fields of study such as Bio-information[4], Image processing[5], Feature selection analysis[6].

A gifted mathematician, Paul Erdos, had over 500 co-authors and published over1400 technical research papers. Thus, the amazing large co-author network of Erdoscan be a good dataset to study.According to the similarity of the link structure of the Web pages and that of thecitation of academic publications, an algorithm inspired from the famous PageRank[7]is proposed to solve the problem. Since we need to work out some proper models andsolutions, the parts below are taken into consideration:(1)Build the co-author network of Erdos1 authors and analyze the properties of the network. (2)Develop a measurement to determine the influence of the co-authors. And then find out the researchers who have significant influence within the network. (3)Build the citation network as to measure the influence of a research paper. (4)Analyze the application of the method we use and apply it on a totally different field. (5)Discuss the proposed method from understanding to utility.

II.RELATED WORK AND OUR ALGORITHM

Our proposed model is based on the following assumptions:(1)The data in Erdos1.htm is complete and reliable .(2)To determine one co-author's influence and impact, we only consider the area given based on the data, although the co-author may has researchers in other areas. (3)The co-authorships in Erdos1.htm can reveal all the efforts of the coauthors in the field.

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 7, July 2014

Nowadays, we are in the era of Internet, we may visit hundreds of web pages per day, The connection among the pages from an extremely huge network. We surprisingly find out that the link structure of the Web is more like the citations of academic publications, as citations and links are essentially similar. However, there are still a number of significant differences between them. For example, the academic papers are scrupulously reviewed and have similar quality and number of citations, while the web pages are extremely uneven. Fortunately, in 1999, an algorithm named PageRank was proposed for rating Web pages[8]. In this paper, we apply PageRank to rank the influence of the co-authors. The network of Web pages is directed, but the cooperations are undirected. To implement the algorithm, we define the edges in the co-author to be bidirectional.

A. The PageRank Algorithm

The original purpose of PageRank is to measure the relative importance of web pages and make a rank. After the algorithm used by Google to order search engine results, it became famous. The applications and modifications of the algorithm immediately became hot topics. In the following, we conduct the mathematical description of PageRank.

As the saying goes, a man is know by the company he keeps. Transferring the knowledge to web site, the more high-quality Web pages referred to a web page, the larger probability of high quality this page may have. The core idea of PageRank is simple but effective. Based on this idea, an intuitive formula can be proposed:

$$R(i)=\sum_{j\in B(i)}R(j) \qquad (1)$$

Where $R(x)$ indicates x 's PageRank and $B(x)$ is the set of pages that point to x. The idea of Formula(1) is that the importance of a page is equal to the the sum of the importance of the webs pointing to it. But there exist a drawback: no matter how many hyperlinks J has, once J points to I, I will get the same importance as J. When J has multiple hyperlinks, this idea can cause non-reasonable condition. For instance, a new website N just have two hyperlinks which point to it. One is from the famous and historical F, while the other comes from an unknown site U. Based on Formula(1), conclusion could be drown that site N gets more importance. This is obviously unilateral. However, we could optimize the formula. When J has multiple hyperlinks(assuming to be N), the importance of each link is obtained as $R(j)/N$. Thus, formula(1) is optimized as:

$$R(i)=\sum_{j\in B(i)}\frac{R(j)}{N(j)} \qquad (2)$$

$N(j)$ is the number of links from j. Figure 1 is cited from Lawrence Page's paper[8] to demonstrate the propagation of rank from one pair of pages to another. From Figure 1, we can obtain the message that, if we want to get the conclusion that N is better than F, we must require that N are able to get many hyperlinks of important web sites or massive hyperlinks of unknown websites which can be accepted. Thus, we consider that Formula(2) expresses the core idea of PageRank algorithm exactly. To obtain standardized results, a constant coefficient C is added to Formula(2). Then we conduct Formula(3):

$$R(i)=C\sum_{j\in B(i)}\frac{R(j)}{N(j)} \qquad (3)$$

Finally, we use a vivid cartoon cited from Wikipedia to illustrate basic principle of PageRank algorithm in Figure 2. The size of each face is proportional to the total size of the other faces which are pointing to it.

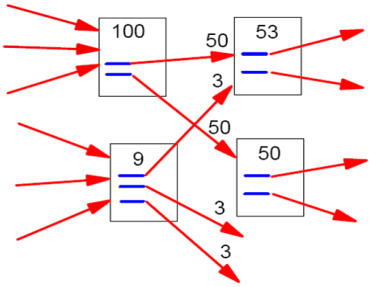

Figure 1. PageRank Calculation

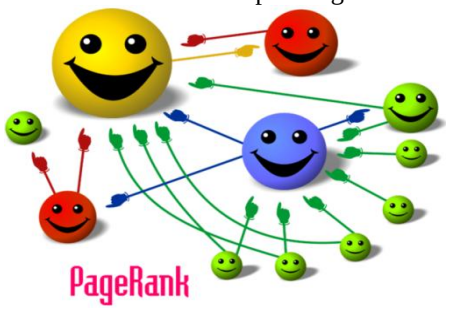

Figure 2. PageRank Illustration

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 7, July 2014

B. The Proposed Novel Algorithm

In order to find out who has significant influence within the co-author network, the network should be built at first. Therefore, data extraction should be conducted to get the structure of the network which is shown in the following section. In addition, we regard 511 co-author as nodes and the co-author as edges. Then, PageRank is applied to deal with the issue.

B.1 Data Extraction

Since there are many useless messages in Erdos1 file, we first clean the data. The coauthors and their links should be extracted. As the quality of PageRank, we just consider the directed connections. That is to say, if two coauthors have direct co-authorship, there is connection between them. Although they may have indirect connection via someone, we don't take this condition into consideration. Fortunately, the data of Erdos1 is given in a normative way which facilitates our work. We can easily find out the co-authors through the years followed by string matching. Then we still use string matching to find out the connections among the co-authors. Finally, we get 511 co-author and 3278 links. Then we use Pajek to draw the structure of network.

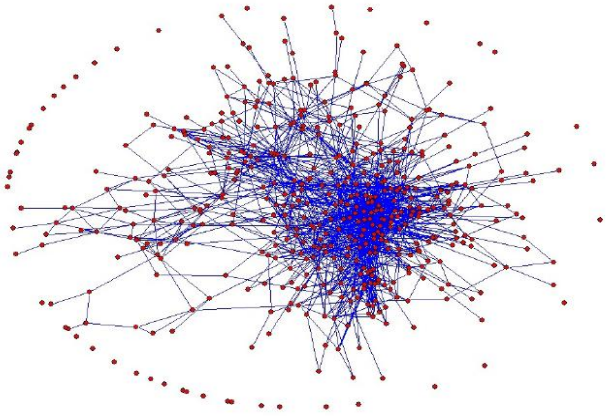

Figure 3. The Co-author Network

In Figure 3, there are some isolated nodes in periphery. This means they haven't cooperated with the rest co-authors. Obviously, these people are not that influential. But the inner nodes are tight for the huge complex connections. In addition, compared to the 511 nodes, the isolated nodes are in small size. In this way, the nodes which have few links can also have the chance to be influential through the network.

B.2 The Influence Measure

After determining the co-author network, the measure of the influence and impact may be the key to the problem. There has been a great deal of work on academic influence and many indexes to measure it have been proposed, such as Science Citation Index(SCI), H-factor, G-factor, A-factor, etc.But the data in this problem is unique. We only have the cooperation information of the co-authors. These measurements above can not fit the problem well. Thus, inspired from the structure of web page network, we introduce the measure of PageRank because this algorithm can give a rank based on the PageRank value. Furthermore, PageRank is brought out by taking advantage of the link structure of Web to produce a global "importance" ranking of every page. The link structure and the cooperations structure are the same in a great degree. Thus, the application of PageRank to this problem is suitable. Next, we apply the algorithm and model to detecting co-authors who have significant influence within the network. The result is shown in the following table:

Table 1. PageRank Rank list

RANK	CO-AUTHOR	VALUE
1	ALON,NOGA	24.07
2	GRAHAM	20.29
3	VOJTECH	19.91
4	BOLLOBAS	19.72
5	HARARY	19.47
6	SOS,VERA	18.75

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 7, July 2014

B.3 Sensitivity and Stability Analysis

In order to test the sensitivity and stability, we select a co-author, BOLLOBAS, BELA, from the network, whose value of PageRank is high. As the influence of BOLLOBAS,BELA is high, we can test whether addition of a researcher who has cooperated with BOLLOBAS,BELA for only one time can confuse the whole network. From the co-authors of BOLLOBAS,BELA, Zhang,Ke Min^1 has 4 times cooperations with other members in the network, which is a suitable condition for test. Assuming that Zhang, Ke Min has a cooperation with Erd□os by way of introduction. Thus, Zhang, Ke Min is added to the network. A short view of the result of the altered network is listed in Table 2.

Table 2. PageRank Rank list

RANK	CO-AUTHOR	VALUE
1	ALON,NOGA	24.00
2	GRAHAM	20.24
3	BOLLOBAS	20.10
4	VOJTECH	19.87
5	HARARY	19.41
6	SOS,VERA	18.71

To described in detail, the rank of Zhang, Ke Min is 249, which is in the middle. But the the result doesn't change much. The fluctuate rate of the nodes is about 0.2%. Most nodes' rank don't change or just change one. But the members cooperated with Zhang change a lot for the The fluctuate rate rate of the PageRank is about 6.5%. Based on the original ranking, the float in turn is 1,2,3,8 and 26. The node ranked latter is a affected larger.

III.THE SIGNIFICANCE OF RESEARCH PAPER

The analysis of co-authorships is a way to measure a scientist's influence. But anotherimportant aspect is the published research papers.In the academic field, most of theresearchers care more about the citations of their papers for the reason that it could reveal the real value of a research paper. For the reason that the citations are from all the academic fields while the co-authorships is only in a small area, we shift our attention from co-authorships to the research papers.

The Citation Network

In this section, we propose the citation network model based on the academic research papers. In the network, we define the papers as nodes and the citation as edges. According to the references relationships between papers, the network comes into being. As the previous co-author network, we still try to apply the PageRank algorithm.

But as discussed before, there are significant differences between web pages and academic publications. The main obstacle is the difference of the structures. The circular will not appear in papers' citations. That is to say, the condition that two papers have mutual citations can't appear. If A cites B, A must be published later than B. But on web pages, there exits cross reference. Co-author network also has cross connections. This phenomenon causes a big problem. To some degree, the citation network is a feedforward network. The structure is of no benefit to PageRank. In this view, the results obtained by the classical PageRank algorithm tend to be the papers published long ago. This is obviously unfair the papers that are published later but have potential academic influence. In order to eliminate the phenomenon, the model or the algorithm must be improved to fit this problem.

A. Our Proposed Method

The model and the algorithm are improved to adapt to the model. As for the core idea of our algorithm, we present a schematic drawing that illustrates the algorithm for the initial understanding ahead.

We still choose the papers in NetSciFoundation.pdf as the set of foundational papers. And the nodes of these papers are what we care. We also give a rank of these foundational nodes. Our work is introducing some other papers which have cited the foundational nodes. These papers are viewed as nodes in the network, but the calculation is different. Furthermore, the improvement is for the purpose of generating cross connection among foundational nodes. The structure is shown in the figure 5. As the figure shows, though the connections of foundational nodes are limited, there are cite nodes in the network connecting several foundational nodes, which may be the bridges for the foundational nodes. As the structure is obvious, we talk about the function of cite nodes and the method we use to improve the

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 7, July 2014

model. For better understanding, we combine figures for detailed instructions. Firstly, we take a foundational node R_i as an example. The Partial view of R_i is given in Figure 6, followed with symbol definition table.

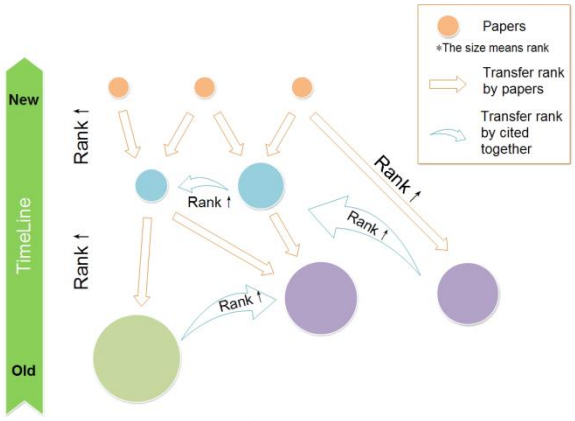

Figure 4. The Core Idea of Proposed Algorithm

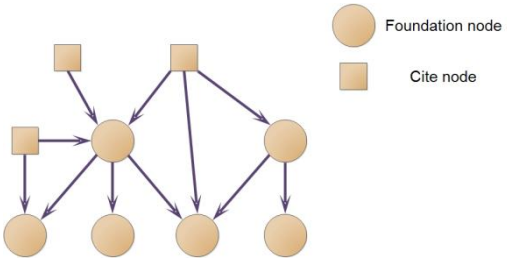

Figure 5. Network with Assistant Nodes

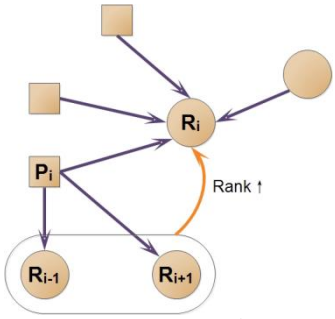

Figure 6. The Value Transfer

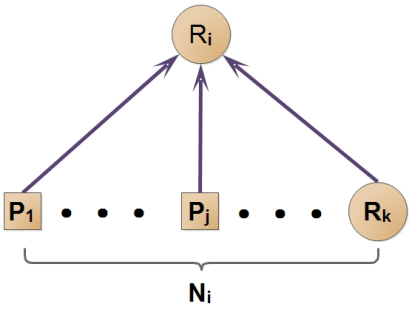

Figure 7. The Definition of N

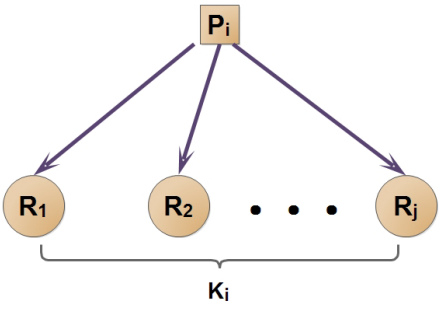

Figure 8. The Definition of K

Table 3. Definition of Symbols

Symbol	Definition
R_i	The foundational node
N_i	Times that R_i has been cited
P_i	The cite node
R_{ij}	The j -th cite nodes of R_i
C_i	Times that P_i has been cited
$\frac{k_i}{R_{ij}}$	Average PageRank of foundational nodes (except R_i) that connect to cite R_{ij}
T_i	Set of nodes cited R_i
M_i	Set of nodes cited by P_i

Then we propose the method in our work and get the iterative formula below:

$$S_i = \sum_{j \in T_i} \frac{\left(\overline{R}_{ij} - R_i\right) \cdot \frac{\ln C_i}{k_i}}{N_i} \tag{4}$$

$$\overline{R}_{ij} = \frac{\sum_{j \in M_i, j \neq i} R_j}{k_i - 1} \tag{5}$$

$$R_i = \begin{cases} R_i + S_i, & if\ (S_i > 0); \\ R_i, & if\ (S_i \leq 0); \end{cases} \tag{6}$$

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 7, July 2014

From the formulas above, S_i reveals the indirect connections of other foundational nodes through the communal cite nodes. In Formula 4, S_i is proportional to $\ln C_i$ and inversely proportional to k_i. This reflects the fact that if a paper is cited by another paper which is cited by a huge number of works, there is high chance that the paper is more influential. In addition, as C_i spans from 100 to 21000, which is too large, we add the "Ln"to narrow the range. To overcome the phenomenon of abusing citing, we get the idea from PageRank to introduce k_i in the denominator. During our work, we find out that if we use original PageRank only, we may get bad astringency. Inspired from the idea of damping factor from the paper|The Anatomy of a Large- scale Hypertextual Web Search Engine Computer Networks and ISDN Systems[9] , we reform Formula(2) into the formula followed.

$$R(i) = \sum_{j \in B(i)} \frac{R(j)}{N(j)} \cdot q + 1 - q \tag{7}$$

In addition, the idea originally comes up to solve the problem of the isolate web pages, which don't cite any other pages. In our work, the isolated nodes also exist. It is joyful to find that introducing damping factor works well both in the isolated nodes and in the astringency after the operation with the improved method. The algorithm converges fast. Figure 9 shows the function clearly:

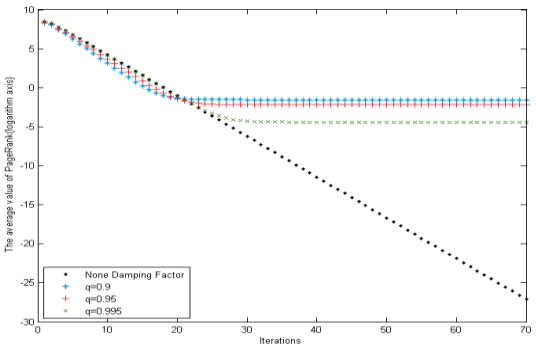

Figure 9. The Function of Damping Factor in Astringency

B. Data Collection

Since the foundational papers are website classical, the citation may be a huge number. It may be a hard work to collect thousands of citation information in a simple way. At first, the web crawler comes to our mind. But, later, we find out that Google seems not friendly to web crawler. Thus, after taking all factors into consideration, a Semi-automatic crawlers alike program, which needs human intervention, comes up. We finally get 3525 lines of raw data. After throwing repeated papers, we get 822 papers(including both the foundational papers and cite papers). That is to say, the total nodes in the network is 822. Firstly, we list the 15 foundational papers and times having been cited. Based on the collected data, we design the network in the following map:

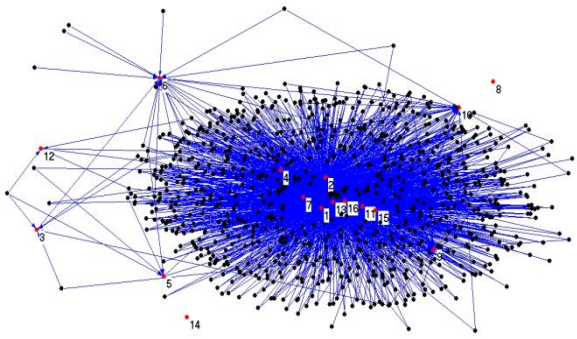

Figure 10. The Structure of Citation Network

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 7, July 2014

C. Results and Analysis

The improved PageRank value of the influence of the papers is listed in Table 4. Furthermore, the first paper has the largest value of RankPage. That is to say, it has the biggest influence. To verify our conclusion, we find out that the first paper is "Collective dynamics of 'small-world' networks"which has been cited for 21688 times. Then we search "Complex network" in Wikipedia, we obtain the message that Two well-known and much studied classes of complex networks are scale-free networks and small-world networks[10]. And in 1998, the paper "Collective dynamics of 'small-world' networks" published in Science by Duncan J.Watts and Steven Strogatz marks the establishment of the small world network model. This information verifies the great influence of the paper. From the work above, we can see the PageRank algorithm indeed has good quality and generalization capabilities. To test our method's ability of application, we implement our method on American Airlines in the next section.

IV.CONCLUSION AND DISCUSSION

A. Further Applications

From the above description, the good generalization capability of our method has been demonstrated. Because our method is operating on the links in the network, if the relationships in the real world can build a network, our method can be a choice to find out the importance or influence rank of all the nodes. If the impact degree of all the nodes in the network is known, the utility is various. For instance, also in the academic field, we can approach the best researchers and try to cooperate with them to quickly increase academic impact. Or we can choose a famous school or thesis advisor, we can do better in our study. Similarly, we will also have more chance to earn both money and good fame if we choose the right business partner. In addition, in other fields, they can change the distribution of the importance of nodes simply by changing the links in the network. At last, we conduct a test to demonstrate the utility of our model and algorithm.

B. Final Conclusion

Determining the influence of academic research is an important work in academy. To build the co-author or citation network is one of the good ways to do the measurement. In this paper, we build co-author network and the citation network respectively. Our algorithm is based on the famous PageRank. According to unique condition in citation network, we make the improvement of both the model and the algorithm to fit the unique problem. Compared with the passenger capacity rank of airports in America gotten from Wikipedia, the implementation shows excellent results. All above reveals the high quality of accuracy and generalization capability. The utility is obvious, and we hope our method can be applied in practice to help making wise decisions. In the future, we would like to use some mathematical tools such as inequations[11] to optimize our proposed method to get better result.

REFERENCES

[1] Impact factor.Retrieved February 11.from http://en.wikipedia.org/wiki/Impact factor
[2] Science Citation Index.Retrieved February 11.from http://en.wikipedia.org/wiki/Science Citation Index.
[3] H-factor.Retrieved February 11.from http://en.wikipedia.org/wiki/H-factor
[4] Yuan Y, Xu Y, Xu J, et al. Predicting the lethal phenotype of the knockout mouse by integrating comprehensive genomic data[J]. Bioinformatics, 2012, 28(9): 1246-1252.
[5] Wang H, Shkjezi F, Hoxha E. Distance metric learning for multi-camera people matching[C]//Advanced Computational Intelligence (ICACI), 2013 Sixth International Conference on. IEEE, 2013: 140-143.
[6] H. Fei and J. Huan, Structured Feature Selection and Task Relationship Inference for Multi-TaskLearning, in Proceedings of the IEEE International Conference on Data Mining (ICDM'11), 2011.
[7] PageRank.Retrieved February 11.from http://en.wikipedia.org/wiki/PageRank
[8] Page L, Brin S, Motwani R, et al. The PageRank citation ranking: bringing orderto the web[J]. 1999.
[9] Brin S, Page L. The anatomy of a large-scale hypertextual Web search engine[J]. Computer networks and ISDN systems, 1998, 30(1): 107-117.
[10] Complex network.Retrieved February 11.from http://en.wikipedia.org/wiki/Complex network
[11] Xu B, Wang X H, Wei W, et al. On reverse Hilbert-type inequalities[J]. Journal of Inequalities and Applications, 2014, 2014(1): 198.