PageRank and analysis of citation cycles

Petr Heller, Michal Nykl, Karel Ježek
Department of Comp. Sc. & Eng.
Fac. of Appl. Sc., University of West Bohemia in Pilsen
Univerzitni 8, 30614 Plzen

Abstract. *The PageRank formula, originally designed for evaluation of significance of web pages, is also usable for evaluation of significance of authors in citation graphs. None of works on this theme takes in notice influence of citation cycles. We are convinced, they should affect the significance of the authors. We used Kendall and Spearman coefficients to evaluate an influence of cycles.*

1 Introduction

PageRank algorithm, originally designed to evaluate the significance of nodes in the hypertext, is used in scientometrics for evaluation the network of the each others quoting authors in author citation graph. The basic formula for calculating PageRank [1] [4] is as follows (modified into an iterative formula):

$$P_{x+1}(A) = \frac{1-d}{n} + d \sum_{u \in B_A} \frac{P_x(u)}{N_u} \qquad (1)$$

where d is damping factor, $Px(A)$ is Pagerank of node,n is count of nodes, u is node links to node A, Nu is count of outgoing links from node u, B_A is set of all edges linked to node A. If we consider the possibility to set the weight of edges in the graph, we get through that a powerful tool for assessing the importance of the individual citations. To analyze a graph with labeled edges, we use the principle of modification known as weighted Pagerank, eg [2] [3].

2 The option of penalizing edges

Under the term "cycle analysis" is understood the detecting of mutual (partner) linking among the authors. So we penalize any mutual agreement between two authors in the manner of "*I will link to you, you will turn me on.*"

We introduce the weight of the edge w is calculated according to the formula (2), where $1/l$ is the *handicap*.

$$w = 1 - 1/l \qquad (2)$$

2.1 Penalizing of the cycles of different lengths simultaneously

The problem occurs when we consider penalizing of different lengths simultaneously. Solving options can be found a few, such as a simple algorithm, where n represents the length of cycle:

- *0)* n is initialized; it's value which represents the length of the longest sought cycle and all the edges are weighted 1,
- *1)* find all cycles of the length n and set the edge in this cycle on the value of $w_0 = 1-(1/n)$,
- *2)* find all cycles of the length n-1 and set the edge in this cycle on the value of $w_1=1-(1/n-1)$,
- ...
- *(N-3)* find all cycles of length 2 and set the edge on the value of $w_{N-2} = 0{,}5$.

We come out from the length n of the cycles, in which the edges are penalized at least. Penalizing of the edges increases gradually in the shorter cycles and the edges weight thus decreases.

The weights of the edges, which were already once set in found cycles, may be overwritten by another value of the penalizing (ie, the edge will be weighted with lower weight). Everything is virtually illustrated in *Fig. 1.*, which represents an author citation graph.

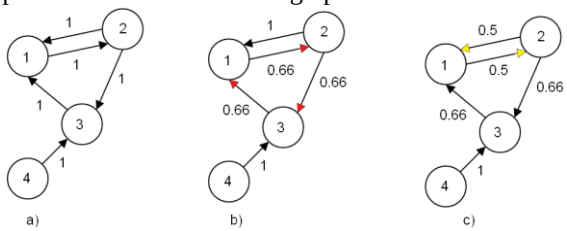

Fig. 1. - Course of edges penalizing in cycles of length three and then two

Fig. 1. represents edges penalizing in the author citation graph for cycles of length three and two. In Fig. 1a) all edges are set to 1 (no penalization). The first step is to find all cycles of length three (Figure 1b)), in the second step all cycles of length two (Figure 1c)). The figure also illustrates an important characteristic of penalization, namely the longer is the cycle, the penalization is *smaller*. So the algorithm first looks for the longest cycle (it's penalization of edges is small) and penalizes edges of nested shorter cycles increasingly.

We realize this method of penalization is not quite fair. The oldest citation or simple cycles, should not be penalized. But, in this case, we need take into account the time of publications. This problem will be solved in our next step. The aim of presented work is to proof how much citing cycles influence rating of authors.

3 Authors rank sensitivity on weights of edges

Experiments were run on the author citation graph obtained from the database DBLP (available at http://www.informatik.uni-trier.de/~ley/db/). The number of authors, publications and citations in the database are summarized in Tab. 1. Tab. 2. summarizes the number of cycles of length 2 and 3 in each individual database.

database	authors	publications	cites (publications)	cites (authors)
DBLP 04	315731	474467	112001	653103
DBLP 06	476564	776401	112262	655232
DBLP 09	749639	2005601	112120	652481

Tab. 1. - Basic statistics DBLP

	DBLP 2004
cycles of length two	22422
cycles of length three	15855

Tab. 2. - The number of cycles of lengths 2 and 3 in DBLP databases

First of all we were interested whether and possibly how much different the results of rankings are, if we penalize the cycles of length two in the author citations graph. We started Pagerank up for a total of eleven different weights in nodes and we compared the results with each other using the Spearman and Kendall correlation coefficients. The resulting values of the Spearman and Kendall coefficients are summarized in Tab 3. and 4.

w of edges	0,001	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1
0,001	1	0,998976	0,996813979	0,994032779	0,990884507	0,987595607	0,984193316	0,980777876	0,977345533	0,97395125	0,970600977
0,1	0,998976	1	0,999295904	0,997633777	0,995396733	0,992855364	0,990094048	0,987232297	0,984286717	0,981324743	0,978360817
0,2	0,996814	0,999296	1	0,999494061	0,998235732	0,996519032	0,994477471	0,99224155	0,989853345	0,987387419	0,984869864
0,3	0,994033	0,997634	0,999494061	1	0,999608806	0,99863643	0,997253917	0,995598075	0,993731984	0,991733963	0,989640432
0,4	0,990885	0,995397	0,998235732	0,999608806	1	0,999697383	0,998914081	0,997790187	0,996406144	0,994842772	0,993144528
0,5	0,987596	0,992855	0,996519032	0,99863643	0,999697383	1	0,999751846	0,999108727	0,998163539	0,996999072	0,995666078
0,6	0,984193	0,990094	0,994477471	0,997253917	0,998914081	0,999751846	1	0,999795705	0,999253531	0,998455983	0,997460403
0,7	0,980778	0,987232	0,99224155	0,995598075	0,997790187	0,999108727	0,999795705	1	0,999825227	0,999364298	0,998679405
0,8	0,977346	0,984287	0,989853345	0,993731984	0,996406144	0,998163539	0,999253531	0,999825227	1	0,999851806	0,999456231
0,9	0,973951	0,981325	0,987387419	0,991733963	0,994842772	0,996999072	0,998455983	0,999364298	0,999851806	1	0,999871856
1	0,970601	0,978361	0,984869864	0,989640432	0,993144528	0,995666078	0,997460403	0,998679405	0,999456231	0,999871856	1

Tab. 3. - Spearman's rank correlation coefficient (d = 0.85)

w of edges	0,001	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1
0,001	1	0,984695	0,971089948	0,958983049	0,947982348	0,937977466	0,928756941	0,920231845	0,912275728	0,904820301	0,897813677
0,1	0,984695	1	0,986278456	0,974062957	0,962963116	0,952837127	0,943507489	0,934880838	0,926831478	0,919288545	0,912193744
0,2	0,97109	0,986278	1	0,987760478	0,9766192	0,966450389	0,957079946	0,948412333	0,940321852	0,93274433	0,925599345
0,3	0,958983	0,974063	0,987760478	1	0,988846848	0,978659991	0,969269799	0,960570524	0,952455466	0,944853406	0,937679332
0,4	0,947982	0,962963	0,9766192	0,988846848	1	0,98980309	0,980404547	0,971691777	0,963557049	0,955943946	0,948746403
0,5	0,937977	0,952837	0,966450389	0,978659991	0,98980309	1	0,990600191	0,981880257	0,973732983	0,966108958	0,958901164
0,6	0,928757	0,943507	0,957079946	0,969269799	0,980404547	0,990600191	1	0,991278562	0,983126816	0,975495983	0,968280393
0,7	0,920232	0,934881	0,948412333	0,960570524	0,971691777	0,981880257	0,991278562	1	0,991846196	0,984207606	0,976982992
0,8	0,912276	0,926831	0,940321852	0,952455466	0,963557049	0,973732983	0,983126816	0,991846196	1	0,992359431	0,985131968
0,9	0,90482	0,919289	0,93274433	0,944853406	0,955943946	0,966108958	0,975495983	0,984207606	0,992359431	1	0,9927685
1	0,897814	0,912194	0,925599345	0,937679332	0,948746403	0,958901164	0,968280393	0,976982992	0,985131968	0,9927685	1

Tab. 4. - Kendall tau rank correlation coefficient (d=0,85)

The collected data serve for easy derivation of the significance of differences between the cases when we will not penalize cycles (edges in the cycle will then have the value of "1") and, if we will. We are interested in how much different the values of the Spearman's and Kendal's criteria are. The size of differences is in the graphs in Fig. 2. and in Fig. 3. The values in the graphs always show the numerical differences between two values of Spearman's and Kendal's coefficients. While in the first graph we are interested in the size of differences between the ranking created by the PageRank without penalizations (original) and ranking with different penalizations (i.e., we compare the original ranking with various penalized rankings) and in the second graph we can see individual increasing between two neighboring variants of rankings (e.g. between rankings created with the weights setting of "0.2" and "0.3").

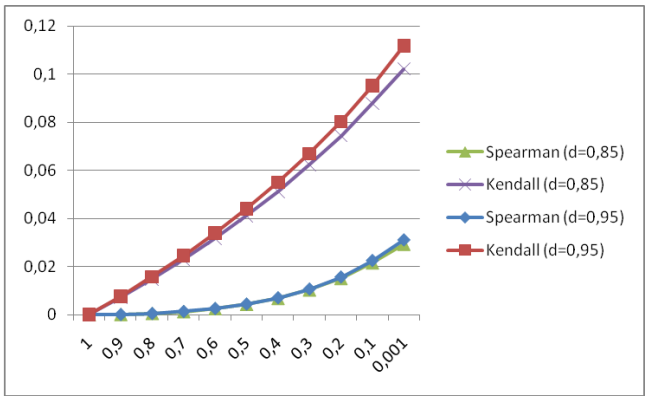

Fig. 2. The size of differences (the case without penalization)

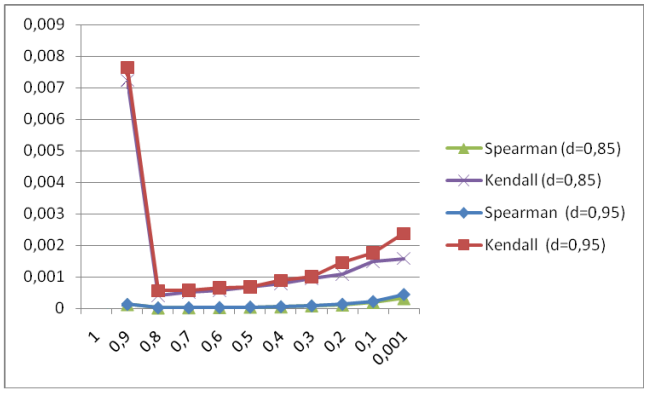

Fig. 3. The size of differences (the case with penalization)

Both rankings show very well that the linear reduction of weights of edges in the cycle causes an almost exponential increase in the diversity rankings.

4 Conclusion

In the paper, some possible solutions of citation cycles were outlined. Experiments with a penalization of cycles were executed and we found that the difference of ranking grows exponentially with increasing penalization. The examination involving the time of each individual citation will be subject of further work. We believe it brings more fairness into results.

5 References

[1] S. Brin, L. Page, R. Motwami, T. Winograd: The PageRank Citation Ranking: Bringing Order to the Web. Technical Report 1999-0120, Computer Science Department, Stanford University, 1999

[2] Wenpu Xing, Ali Ghorbani. Weighted PageRank Algorithm. Proceedings of the Second Annual Conference on Communication Networks and Services search, 2004

[3] Karel Jezek, Dalibor Fiala, Josef Steinberger: Exploration and Evaluation of Citation Network. 2007. ELPUB 2007 Conference of Electronic Publishing Toronto

[4] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems, 30(1–7):107–117, 1998.