Selection of Suitable PageRank Calculation for Analysis of Differences between Expected and Observed Probability of Accesses to Web Pages

Jozef Kapusta[1][0000-0002-8285-2404], Michal Munk[2][0000-0002-9913-3596] and Peter Svec[3][0000-0002-1713-6444]

[1] Constantine the Philosopher University in Nitra, Slovakia, jkapusta@ukf.sk
[2] Constantine the Philosopher University in Nitra, Slovakia, mmunk@ukf.sk
[3] Constantine the Philosopher University in Nitra, Slovakia, psvec@ukf.sk

Abstract. We describe various approaches how to calculate the value of PageRank in this paper. There are few methods how to calculate the PageRank, from the basic historical one to more enhanced versions. Most of them are using the original value of the damping factor. We describe the experiment we realised using our method for analysing differences between expected and observed probability of accesses to web pages of the selected portal. We used five slightly different methods for PageRank estimation using both the original value of damping factor and the value calculated from data in the web server log file. We assumed and confirmed that the estimation/calculation of the damping factor would have a significant impact on the estimation of the PageRank. We also wrongly assumed that the estimation/calculation of the damping factor would have a significant impact on the number of suspicious pages. We also compared the computational complexity of used PageRank methods, and the most effective method seems to be a method with the estimated value of the damping factor.

Keywords: web usage mining, web structure mining, PageRank, damping factor, support, observed visit rate, expected visit rate.

1 Introduction

We can find many web mining methods that try to solve different issues of websites, like employing some personalisation, improve the structure of the website or reorganise web pages itself. Only a few of these methods try to combine the web structure and the web usage mining methods to achieve this aim. We developed method described in [1, 2] to analyse the differences between expected and observed probability of accesses to web pages of the selected portal. The expected rate of access to the web page was estimated using the PageRank (PR); the real visits were gotten from the web server log file. After the data pre-processing and user session identification [3], we calculated the value of support, which represent the real visits to the website. The method of calculating the PR is essential for our experiment. There are many different methods for calculating the PR [4–7], and we try to find the ideal one for our method of finding differences

between expected and observed probability of accesses to web pages. We are also looking for the ideal value of the damping factor, which will be discussed later. We combine various methods of PR calculation with different methods of setting the damping factor value – *d* in this paper.

2 Related Work

PR was developed at Stanford University by Larry Page and Sergey Brin [8, 9]. PR is a simple, robust and reliable way to measure the importance of web pages which can be computed offline using just the structure of web pages (sitemap) and the hyperlinks between pages. The PR form a probability distribution over web pages, so the sum of all web pages' PR will be one. PR can be thought of as a model of user behaviour. The original PR assume that there is a "random surfer" who is given a web page at random and keeps clicking on links, never hitting "back" but eventually gets bored and starts on another random page. The probability that the random surfer visits a page is its PR. Moreover, the damping factor is the probability at each page the "random surfer" will get bored and request another random page. The usual value of the damping factor is 0.85.

The literature does not offer the best value of the damping factor. When damping factor gets close to 1, the Markov process is closer to the "ideal" one, which would somehow suggest that damping factor should be chosen as close to 1 as possible. Boldi, Santini and Vigna [10] give several proofs of what happened when we choose the wrong value of the damping factor. When the value of *d* is close to 1, many important web pages will have rank 0 in the limit. Choosing *d* too close to 1 does not provide any good PR. Rather, PR becomes "sensible" somewhere in between 0 and 1.

The simplest and most basic algorithm that computes PR is an application of the Power Method. The PR algorithm including the crawler is in detail described in [11].

Some researchers try to use another approaches or enhancements to the PR. Deore and Paikrao [12] describe the UsersRank algorithm. While browsing the web, a user can save the link as a personal bookmark or as a social one. The social bookmark is shared among multiple users. UsersRank algorithm makes use of these bookmarks and produces valuable information for search engines. It believes in the logic that if the user is having some links as bookmarked, then those links are used by someone hence really valuable and gives useful results for web searches. Every bookmarked entry is considered as a vote given by the user to that page. UsersRank is achieved by summing up a total number of votes given by the users to that page.

Wang and Tao [13] create Personalised PR and combine the Monte-Carlo approach and group target nodes with similar PR together. They introduce a new notion called "PageRank heavy hitter" to quantify the importance of the second direction, and thereby, gives a convenient way to harness this direction for the recommendation. Personalised PR have been widely applied in social networking services to make friend recommendations; this is usually done by leveraging only the first "direction of importance".

The PR can be manipulated by the community of webmasters. They can create good links between pages and raise the rank. Yang, King and Lyu [14] tried to handle the manipulation problem, and they offer a DiffusionRank algorithm. This rank is motivated by the heat diffusion phenomena, which can be connected to web ranking because the activities flow on the web can be imagined as heat flow. They propose that link from a page to another can be treated as the pipe of an air-conditioner, and heat flow can embody the structure of the underlying web graph. Even in this idea, the authors used the value of 0.85 for the damping factor.

Yoseff and Mashiach [15] use Reverse PR based on the reverse graph (obtained by reversing the directions of all links) which is more suitable for local PR approximation and admits fast PR convergence.

Eiron, McCurley and Tomlin [16] propose some innovations (HostRank and DirRank), to detect also pages that cannot be reached by crawlers. They call those pages as frontiers, and the consider them as significant – original PR algorithm deletes dangling pages. Their experiment was done on major newspapers in the U.S., and again they used the value of 0.85.

In our previous experiments [1, 2], we proved that there is a higher dependence of PR on the value of *support* in the visit rate of the examined web pages when the log file with identified user sessions is well-prepared. We also proved that the expected visit-rate of the individual web page (variable *PR*) correlated with the real visit-rate (*support*) obtained from the web server log file using the web usage mining method. We also proved that the dependence of PR on variable *support* would be higher when we preprocess the log file using user session identification methods. We utilized the potential advantages of joining web structure and the web usage mining methods in the residual analysis.

3 Materials and Methods

We developed a basic crawler, which went through and analysed web pages. The crawler created a sitemap which we have utilized later in the PR calculation of individual pages. The crawler was simple because it scanned only the hyperlinks between web pages. We consider this as the main limitation of the proposed method because the crawler did not regard the actual position of the hyperlinks within the web page layout, which has a strong influence on the probability of being accessed by a website visitor.

Consequently, we calculated PR for different web pages, based on several version of PR. We consider PR as a static evaluation of web pages, i.e. PR is calculated for each web page off-line, independently of search queries.

We divide the web page hyperlinks into two categories:

- In-links – all hyperlinks that refer to the web page *i* from other web pages.
- Out-links – all hyperlinks that refer to other web pages from the web page *i*.

The recursive hyperlinks are not considered. At the same time, we assume, the hyperlink from the web page, which referred to other web pages, transferred its importance to the target web pages implicitly. It means the web page is more relevant if other important web pages refer to it. We consider the web as an oriented graph *G*

$$G = (V, E), \tag{1}$$

where *V* is a set of nodes, i.e. a set of all web pages and *E* is a set of oriented edges, i.e. hyperlinks among web pages.

Let *n* (*n*=|*V*|) be the total count of web pages of the website. Then the PR of the web page *i* (p_i) is defined as

$$p_i(0) = \frac{1}{n}, \tag{2}$$

$$p_i(t+1) = \sum_{(j,i)\in E} \frac{p_j(t)}{o_j}, j \in V, \tag{3}$$

where o_j is the count of hyperlinks referred (Out-links) to the other web pages from the web page *j*. This is the first, most simple version of the PR algorithm. We also used improved versions, employing the Random Surfer Model using the damping factor *d* using both available versions of calculating the PR. The first improved version is as follows

$$p_i(t+1) = 1 - d + d \sum_{(j,i)\in E} \frac{p_j(t)}{o_j}, j \in V, \tag{4}$$

where the value of *d* fits the interval ⟨0, 1). The most common value of *d* is 0.85.

The second improved version of PR is the damping factor subtracted from 1 is divided by the number of pages, so the sum of PRs is equal to 1

$$p_i(t+1) = \frac{1-d}{n} + d \sum_{(j,i)\in E} \frac{p_j(t)}{o_j}, j \in V. \tag{5}$$

We can iterate this calculation until the difference between the two following values of *Pr(i)* will be less than the desired accuracy *ε*.

We are calculating the internal PR, which is bounded by the domain of the website. We are interested only in links of the same portal. We are comparing different methods of calculating PR, including different values of damping factor in our methods of the analysis of differences between expected and observed probability of accesses to web pages

3.1 Analysis of differences between expected and observed probability of accesses to web pages

We assume that the website is an oriented graph as stated in (1). We can identify suspicious pages using the following sequence:

1. **calculating the observed probability *(s)*** of accesses to web pages $i \in V$

$$s_i = \frac{number\ of\ identified\ sequences\ with\ i}{total\ number\ of\ identified\ sequences}, \tag{6}$$

2. **estimation of expected probability** of access to web pages $i \in V$ with each method for calculating the PR as stated in (3), (4), (5).
3. **visualization of difference between the expected *(p)* and observed *(s)* probability**

$$r_i = s_i - p_i, \tag{7}$$

4. **identification of suspicious pages**

$$\bar{r} \pm 2s. \tag{8}$$

The analysis of the expected visit and observed visit combine data sources for web usage analysis and web structure analysis. Evaluating the structure of the web means to identify suspicious sites. Suspicious pages are pages where expected visit do not match to the observed one. We use the visualization of differences in observed and expected access probabilities (7) and the identification of extreme differences (8) for the evaluation. Observed page access probabilities are represented by the level of *support* for the frequent one-element item sets (6) and the expected access probabilities are represented by the PR (3), (4), (5). The PR for a particular page reflects the likelihood a random visitor will get to this page. While the observed access probability - *support* is calculated from the pre-processed portal web server log file, the PR for the examined web portal is calculated from the sitemap.

The PR method is based on the principle - the better the page is, the more links point to it. The PR value of page *i* depends on the extent to which is the recommender is important (*pj*) and how much recommendation it gives (*oj*). In other words, the PR value of one page depends on the PR of the referral page and the number of links it refers to (2). The value of *t* is the iteration number, given that PR is counted recursively. In the iteration process, all pages start with the same PR (2). If the page does not contain a link, e.g. document or image, then we assume that the user will go to any page, i.e. as if it contained *n* links (to all the pages of the examined portal).

4 Research methodology

The value of damping factor *d* is significant in our experiment when using calculations (4) and (5). It indicates the probability that a random visitor comes to a page directly (not from a link). We have records from the web server log file in our experiment. The

log file contains referrer information for each access, so we know where the visitor came from. We can estimate the value of the damping factor $\hat{d}$ as the proportion of pages with a referrer within the examined portal *P*.

$$\hat{d} = \frac{the\ number\ of\ accesses\ with\ referer\ i}{total\ number\ of\ access\ to\ the\ portal\ P}, i \in P. \quad (9)$$

Our aim is to compare different approaches to estimate the value of PR for the analysis of differences between expected and observed probability of accesses to web pages and to verify that the estimation of the damping factor has a significant effect on the reliability/accuracy of the expected page access. Using the experiment, we want to verify the following assumptions:

(1) We assume that the estimate/calculation of parameter d will have a significant effect on PageRank estimation.

(2) We assume that the estimate/calculation of parameter d will have a significant impact on the number of identified suspicious pages.

The experiment then consists of the following steps:

1. Determine the observed probability of access to web pages represented by the value of *support*;
2. Estimate of expected probability of access to web pages represented by the estimated value of PageRank:
   a. *PR A* - estimate the value of PR based on (3),
   b. *PR B* - estimate the value of PR based on (4) for d=0.85,
   c. *PR C* - estimate the value of PR based on (4) with an estimated value of $\hat{d}$,
   d. *PR D* - estimate the value of PR based on (5) for d=0.85,
   e. *PR E* - estimate the value of PR based on (5) with an estimated value of $\hat{d}$;
3. Make a linear transformation of results;
4. Identify dependence among examined variables;
5. Compare different value PR with considering the value of *support*.
6. Visualize the differences of observed and expected probabilities represented by different methods for estimation of PR;
7. Identify suspicious pages;
8. Qualitative evaluate identified suspicious sites using various PR methods.

5 Results

The sitemap of examined portal consists of 3996 pages. The damping factor calculated according to (9) is $\hat{d}$ = 0.35. Kendall's coefficient of concordance represents the degree of concordance in values of the residuals using different PR estimations. The value of the coefficient (Table 1) is approximately 0.01 while 1 means a perfect concordance and 0 represents a discordance. Low values of the coefficient confirm statistically significant differences.

Table 1. Homogeneous groups for residuals

residual	mean	1	2
residual PR E	-0.1567	****	
residual PR C	-0.1567	****	
residual PR B	-0.1531	****	****
residual PR D	-0.1531	****	****
residual PR A	-0.1521		****
Kendall coefficient of concordance		0.00989	

Based on multiple comparisons (LSD test) two homogenous groups (Table 1) were identified regarding the average residual for different PR estimations. Statistically significant differences were proved at the 5 % significance level in the average residual between a basic estimation of *PR A* and estimations *PR C* and *PR E*, which were estimated based on calculated parameter $\hat{d}$.

Table 2. Homogeneous groups for residuals considering page level

level	residual	mean	1	2	3
higher	**residual PR E**	-0.1922		****	
higher	**residual PR C**	-0.1922		****	
higher	**residual PR B**	-0.1880		****	****
higher	**residual PR D**	-0.1879		****	****
higher	**residual PR A**	-0.1868			****
lower	**residual PR C**	0.0006	****		
lower	**residual PR E**	0.0006	****		
lower	**residual PR B**	0.0016	****		
lower	**residual PR D**	0.0016	****		
lower	**residual PR A**	0.0019	****		
higher level: Kendall coefficient of concordance				0.01515	
lower level: Kendall coefficient of concordance				0.76562	

A closer look at the results (Table 2) shows that

- A high concordance in residual values, when using different PR estimations is in the case of a lower page level (> 2). The value of the coefficient of concordance (Table 2) is approximately 0.77, i.e. a high concordance. In the case of pages at the lower page level, statistically significant differences were not identified.
- On the contrary, statistically significant differences were identified in the case of pages with a high page level (< 3). The value of the coefficient of concordance (Table 2) is approximately 0.02, i.e. discordance. Statistically significant differences

were proved at the 5% significance level in the average residual between a basic estimation of *PR A* and estimations *PR C* and *PR E*.

Table 3. Correlations: support & PageRank

	valid N	r	t	p-value
support & PR A	174	0.3196	4.4242	0.000017
support & PR B	174	0.3170	4.3837	0.000020
support & PR C	174	0.3204	4.4356	0.000016
support & PR D	174	0.3170	4.3829	0.000020
support & PR E	174	0.3205	4.4371	0.000016

Between the *support* measure and *PR* estimations (Table 3) was identified a moderate measure of direct proportional of dependency. The correlation coefficients for all PR estimations (Table 3) are statistically significant, with a slightly higher dependency between *support* and PR estimations *PR C* and *PR E*. In all cases, *PR* values (expected visit rate) and *support* values (observed visit rate) are changed together in the same direction, where the highest positive correlations were reached for *PR* estimates with the suggested *damping factor* estimate.

Table 4. Kendall tau correlations for PageRank estimations at a high page level

level = higher	PR A	PR B	PR C	PR D	PR E
PR A	1.0000	0.8851	0.7672	0.8853	0.7680
PR B	0.8851	1.0000	0.8820	0.9998	0.8829
PR C	0.7672	0.8820	1.0000	0.8818	0.9992
PR D	0.8853	0.9998	0.8818	1.0000	0.8827
PR E	0.7680	0.8829	0.9992	0.8827	1.0000

Table 5. Kendall tau correlations for PageRank estimations at a lower page level

level = lower	PR A	PR B	PR C	PR D	PR E
PR A	1.0000	0.7686	0.5764	0.7686	0.5764
PR B	0.7686	1.0000	0.8079	1.0000	0.8079
PR C	0.5764	0.8079	1.0000	0.8079	1.0000
PR D	0.7686	1.0000	0.8079	1.0000	0.8079
PR E	0.5764	0.8079	1.0000	0.8079	1.0000

The lowest measure of concordance was identified between a basic PR estimation *PR A* and estimations *PR C* and *PR E* for a high page level (< 0.77) as well as for a lower page level (< 0.58).

The first assumption was confirmed, the estimate/calculation of parameter *d* has a significant effect on PR estimation. PR estimations *PR C* and *PR E* provided the most accurate results - the highest degree of concordance was achieved with the variable *support*. Moreover, statistically significant differences in the values of residual were proved between a basic PR estimation *PR A* and estimations *PR C* and *PR E*, the highest differences being shown for a higher page level (< 3). Similarly, the lowest measure of concordance was identified in values of PR between a basic PR estimation A and estimations *PR C* and *PR E*, which were estimated based on the calculated parameter *d*.

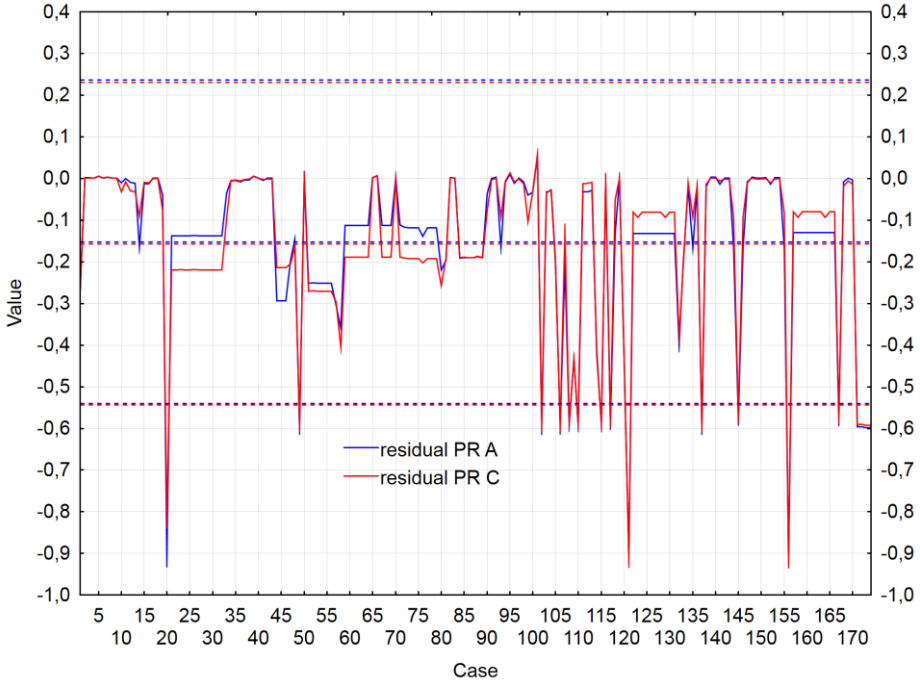

Fig. 1. Identification of suspicious pages based on the 2sigma rule

Figure 1 visualizes the differences between observed and expected probabilities of accesses of the web users, represented by the measure *support* and PR estimations *PR A* and *PR C*. Larger differences occurred in pages with a higher PR. After applying the "2 * standard deviation" rule, we identified 17 extreme cases- suspicious pages. In all cases, the expectations of the page creators were overestimated in terms of visit rate. Specifically, there are pages of level 1, which were characterized by a high PR (based on all examined PR estimations *PR A* - *PR E*) and the low observed visit rate (*support* < 2.2%).

The second assumption was not confirmed, estimation of parameter d does not have a significant impact on the number of identified suspicious pages. Regardless of the

used PR estimation, for the representation of the expected visit rate, the same suspicious pages were identified, i.e. the pages where the expectations of the web creators about the visit rate were overestimated.

6 Discussion

We realized experiment to verify the appropriateness of different methods for calculating the PR in the method of the analysis of differences between expected and observed probability of accesses to web pages. Before realizing the experiment, we defined two research assumptions. We assumed that the estimation/calculation of damping factor *d* would have a significant impact on the estimation of the PR. This assumption has been confirmed, the estimation/calculation of parameter *d* has a significant effect on PR estimation. Estimation of *PR C* and *PR E* have produced the most accurate results – we achieved the highest degree of correlation with the variable *support*. There were statistically significant differences in the residual value among the estimation of *PR A*, estimation of *PR C*, and estimation of *PR E* estimates. The largest differences were for a higher level of pages (a level lower than 3, the main page is level 0). Similarly, the lowest level of correlation was identified among basic *PR A*, *PR C*, and *PR E*; all are employing estimated value of damping factor.

The second assumption that the estimation/calculation of the damping factor will have a significant impact on the number of suspicious pages has not been confirmed. The estimation of the damping factor *d* does not have a significant impact on the number of suspicious pages. The same pages were identified as suspicious pages regardless of the method used for calculating the PR of expected access. Suspicious pages were overestimated pages from the webmaster point of view.

Another important factor for calculating the PR is the computational complexity. We can see in Table 6 the number of iterations needed for each page rank method we used. All calculations were made with the accuracy of 0.00005.

Table 6. The computational complexity of the different methods of calculating PR

Method	Required accuracy	Number of iterations needed
PR A	0.000005	121
PR B	0.000005	74
PR C	0.000005	18
PR D	0.000005	39
PR E	0.000005	10

The most effective method seems to be *PR E* and *PR C* both with the estimated value of the damping factor.

7 Conclusion

The different web mining methods and techniques can help to solve some typical issues of the contemporary websites, contribute to more effective personalization, improve a website structure and reorganize its web pages. The analysis of differences between expected and observed probability of accesses to web pages can give a hint if and how the combination of web structure mining method and web usage mining methods can identify misplaced pages and how they can contribute to the improvement of the website structure. The method analyses the relationship between the estimated importance of the web page from the webmaster point of view using the web structure mining method based on PR and visitors' real perception of the importance of that individual web page. The method compares the real access from the web server log file the estimated accesses using the PR algorithm. There are several options for calculation of PR. We compared these methods and proposed own modification of the PR algorithm. We employed the estimation of the damping factor and using the experiment we verified that this modification is most appropriate. Our calculated value of damping factor was 0.35 while to most commonly used value is 0.85. We compared the impact of the value of damping factor to PR estimations and methods with calculated damping factor provide the most accurate results with fewer iterations. The problem of the method may be the dynamics of the pages created. In most portals, new sites are growing every day. The PR calculation itself always works with the actual number of pages, i.e. new pages will automatically include in its calculation. However, it takes some time for the new pages of the portal to be visited and accesses will be part of the log file. This may slightly distort the estimate of the dumping factor needed for the calculation.

Acknowledgements

This work was supported by the Slovak Research and Development Agency under the contract No. APVV-14-0336, and Scientific Grant Agency of the Ministry of Education of the Slovak Republic and of Slovak Academy of Sciences under the contracts No. VEGA-1/0776/18, and by the scientific research project of the Czech Sciences Foundation Grant No. GA16-19590S.

References

1. Kapusta J, Munk M, Drlík M (2014) Analysis of Differences between Expected and Observed Probability of Accesses to Web Pages. In: Hwang D, Jung JJ, Nguyen N-T (eds) Comput. Collect. Intell. Technol. Appl. Springer International Publishing, Cham, pp 673–683
2. Kapusta J, Munk M, Drlík M (2015) Identification of Underestimated and Overestimated Web Pages Using PageRank and Web Usage Mining Methods. In: Nguyen NT (ed) Trans. Comput. Collect. Intell. XVIII. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 127–146

3. Srivastava M, Garg R, Mishra PK (2015) Analysis of Data Extraction and Data Cleaning in Web Usage Mining. In: Proc. 2015 Int. Conf. Adv. Res. Comput. Sci. Eng. & Technol. (ICARCSET 2015). ACM, New York, NY, USA, p 13:1--13:6
4. Migallón H, Migallón V, Palomino JA, Penadés J (2018) A heuristic relaxed extrapolated algorithm for accelerating PageRank. Adv Eng Softw 120:88–95. doi: 10.1016/j.advengsoft.2016.01.024
5. Shen Z-L, Huang T-Z, Carpentieri B, et al (2017) An efficient elimination strategy for solving PageRank problems. Appl Math Comput 298:111–122. doi: https://doi.org/10.1016/j.amc.2016.10.031
6. Csáji BC, Jungers RM, Blondel VD (2014) PageRank optimization by edge selection. Discret Appl Math 169:73–87. doi: https://doi.org/10.1016/j.dam.2014.01.007
7. Buzzanca M, Carchiolo V, Longheu A, et al (2018) Black hole metric: Overcoming the pagerank normalization problem. Inf Sci (Ny) 438:58–72. doi: https://doi.org/10.1016/j.ins.2018.01.033
8. Page L, Brin S, Motwani R, Winograd T (1999) The PageRank Citation Ranking: Bringing Order to the Web. Stanford InfoLab
9. Brin S, Page L (1998) The anatomy of a large-scale hypertextual Web search engine. Comput Networks ISDN Syst 30:107–117. doi: https://doi.org/10.1016/S0169-7552(98)00110-X
10. Boldi P, Santini M, Vigna S (2005) PageRank As a Function of the Damping Factor. In: Proc. 14th Int. Conf. World Wide Web. ACM, New York, NY, USA, pp 557–566
11. Benincasa C, Calden A, Hanlon E, et al (2006) Page Rank Algorithm. 21.
12. Deore AD, Paikrao RL (2012) Ranking Based Web Search Algorithms. Int. J. Sci. Res. Publ. 2:
13. Wang S, Tao Y (2018) Efficient Algorithms for Finding Approximate Heavy Hitters in Personalized PageRanks. In: Proc. 2018 Int. Conf. Manag. Data - SIGMOD '18. ACM Press, New York, New York, USA, pp 1113–1127
14. Yang H, King I, Lyu MR (2007) DiffusionRank: A possible penicillin for web spamming. In: Proc. 30th Annu. Int. ACM SIGIR Conf. Res. Dev. Inf. Retr. - SIGIR '07. ACM Press, New York, New York, USA, p 431
15. Bar-Yossef Z, Mashiach L-T (2008) Local Approximation of Pagerank and Reverse Pagerank. In: Proc. 17th ACM Conf. Inf. Knowl. Manag. ACM, New York, NY, USA, pp 279–288
16. Eiron N, McCurley KS, Tomlin JA (2004) Ranking the Web Frontier. In: Proc. 13th Int. Conf. World Wide Web. ACM, New York, NY, USA, pp 309–318