Focused Page Rank in Scientific Papers Ranking

Mikalai Krapivin and Maurizio Marchese

[1] Dipartimento di Ingegneria e Scienza dell' Informazione (DISI), University of Trento, Italy

krapivin@disi.unitn.it, marchese@disi.unitn.it

We propose Focused Page Rank (FPR) algorithm adaptation for the problem of scientific papers ranking. FPR is based on the Focused Surfer model, where the probability to follow the reference in a paper is proportional to its citation count. Evaluation on Citeseer autonomous digital library content showed that proposed model is a tradeoff between traditional citation count and basic Page Rank (PR). In contrast to basic Page Rank, proposed Focused Surfer model suffers less from the "outbound links" problem. We believe that FPR algorithm is closer to reality because highly cited papers are more visible and tend to attract more citations in future. This is in accordance with the one of the most significant principles of Scientometrics. No need for lexical analysis of the domain corpus and simplicity of implementation are among the strong points of the proposed model and make the proposed ranking technique attractive for academia digital libraries.

Keywords: Scientometrics, Page Rank, Focused Surfer, Citation-based metrics, Digital Libraries

1. Introduction

Ten years ago Google corporation applied Page Rank (PR) algorithm [1] with great success to the problem of web-pages ranking. PR algorithm is purely statistical, and there is no need to analyze the content of each page lexically. It uses a "*Random Surfer*" model [1] in which the process of browsing through the web pages links is modeled by the stochastic Markov process, fully described by a Markov chain matrix. Recently Page Rank has been studied from several points of view including computational feasibility, modifications and adaptations to the different types of graphs and network models, probabilistic model, mathematical background [2]. Its popularity for ranking web-pages makes it popular in other domains, like ranking of scholarly publications.

The most intriguing question about PR is how to compute it for the whole web? Whole internet contains terabytes of information, and being represented as a graph it exceeds modern computers' memory. It is a creative engineering task to design fast access storage to compute PR. Let us briefly outline major methods for PR computation.

1) The simplest one is the cyclic PR computation for all nodes – one by one - in the graph, using recursive formula (1) until convergence [3]. This method takes unit vector as initial rank approximation.
2) PR authors, Brin and Page proposed polynomial convergence method [1], similar to Jacobi methods.
3) Method (2) was improved by Haveliwala in 1999 [4] using "block-based strategy", similar to implementations in relational database products.
4) In 2003 Langille [5] invented the procedure with reduction of the iterations number with lucky initial approximation.
5) In 2003 Kamvar et al.[6], proposed quadratic extrapolation method to accelerate PR convergence and evaluated their methodology under roughly 81 millions of pages.

Most of mentioned above works are related to the Web links ranking problem which usually deals with much larger graphs than scientific citing problem. So, the computation problem has been studied well enough and looks feasible.

A correlated research topic is related to promising PageRank modifications, for instance:

1) PR Computation with or without damp factor (see formula (2) below).
2) Personalized Page Rank with some initial personalization vector is more common for web-search engines. Here all pages have their own personal weights *before* PR calculation.
3) Focusing of PR, or redistribution of links to link probabilities in the stochastic Markov matrix. This means that core PR model of *Random Surfer* is no longer Random, it becomes focused. This model was successfully applied to the web pages ranking problem by Tony Abou-Assaleh *et al.* [7] and by Fuyong Yuan *et al.*[8] in 2007.
4) Double (or more) focusing of PR takes into account more deep properties of citation graph entities during stochastic Markov matrix composition. For example, it may first focus on site name and then on site content.

Ranking problem is also very important in the scholarly domain, where the main metrics of an article's contribution is the *citations count* [9]. Recently Chen and others [3] applied Page Rank idea for the scientific citations. The major result of this application is that some classical articles in Physics domain have small quantity of citations and very high Page Rank. Chen *et al.* called them "*scientific gems*". Existence of "scientific gems" is caused by PR model which captures not only the total citation count, but the rank of each of the citing papers.

Another Page Rank adaptation for the same problem was performed by Yan Sun *et al.*, 2007 [10]. They applied "personalization" modification from above, where personalized vector was taken in proportion to the publishing journals weight. Then the validity of the rank was estimated by the *cumulative gain* function [10].

Recently Page Rank was successfully applied to the problem of assessing papers, institutions, authors for really large scale problem (~billion of items) [12]. Both methods of assessing academia papers – the traditional citation count and the more recent PageRank and are based on the quantity of citations. Citation count advantages are I) simplicity of computation; II) it is a proven method which has been used for

many years in scientometrics. Proven history of use is very important in the conservative academia domain. Page Rank has the following strong sides: I) it statistically analyses whole citations graph at once; II) it captures not just quantity, but also *quality* of citing papers. However, Page Rank algorithm introduces also computational artifacts like the “effect of outbound links” [13]: this means that if a paper *P* is cited many times by papers with high rank but containing a large quantity of outgoing links — it may decrease *P*’s rank. Situation when a paper is highly cited but poorly ranked by PR looks strange for academia publications.

In this paper, we propose Focused Page Rank adaptation to reduce the "effect of outbound links" and to make a tradeoff between Page Rank and Citation Count. In our proposed model a “reader of an article”[1] may follow all references with different probabilities, so our random surfer model is getting focused. We take citation count as a measure of attractiveness of a reference inside a scientific paper.

2. Problem statement

Let us briefly outline what the original Page Rank algorithm does. It performs ranking for the nodes of the oriented graph with *N* vertices. There are two different link types which may connect node to the neighbors: outbound links and inbound ones. The main measure of node’s weight is inbound links quantity. When we apply this model to the scientific citations problem, we can establish the following similarities: papers are nodes of the graph; citations made by the other papers are inbound links; "references" section creates outbound links set. This is true for most of scientific papers. Rank of a node according to PR is given by the recursion formula (1):

$$P_i = \sum_{i \neq j}^{j \in D} \frac{P_j}{S(j)}, \tag{1}$$

where *S(j)* is the quantity of references for paper P_j, $i,j \in \{1,..,n\}$ are paper sequence numbers in a graph and D_i is variety of all articles which cite article *i*. In the matrix from we can rewrite it as eigenvector problem (2):

$$\vec{r} = A \cdot \vec{r}, \tag{2}$$

where A is the transition matrix, or stochastic Markov matrix. This consideration exposes several potential problems in rank computation as discussed in [2],[5]. One of them is the presence of the papers which cite other papers but are not cited themselves. They are called dangling nodes and they may be treated as the most recent papers. In this case equation (2) may have no unique solution, or it may have no solution at all. It will lead to zero-rows occurrence in the transition matrix and uncertainty of the rank of dangling nodes. Such problem may be resolved with the introduction of a damp-factor d. The damp (or decay) factor is a positive number d, such that $0 < d < 1$ and we illustrate it in formula (3):

[1] “reader of an article” is a Focused Surfer.

$$P_i = (1-d) \cdot \sum_{i \neq j}^{j \in D} \frac{P_j}{S(j)} + \frac{d}{N} \quad (3)$$

Damp factor was proposed by PR inventors Page and Brin and widely used in different Page Rank computations. It helps to achieve two goals at once: 1) faster convergence using iterative computational methods; 2) problem becomes solvable for sure since all nodes have a possibility to be visited by a Random Surfer.

2.1 Scientific Citations Graph Specific Characteristics

When considering the scientific citation problem we may avoid the mentioned above problems in a very natural way because of the following peculiarities of our specific domain (i.e. scientific papers):

I) After and article is published, it cannot cite anymore.

II) If the number of articles in the graph is *N*, each paper may potentially have from *1* to *N-1* ingoing links and the same quantity of outgoing ones. Since *N>>1,* in real life the citation graph is extremely sparse. Indeed, articles normally have from 5 to 20 citations inside, comparing average quantity of citations per article *m* with quantity of papers in graph *N* it is obvious that *m<<N*.

First condition simplifies highly the problem because citations graph becomes unidirectional. We assume (and experimentally prove) that citation graph is free of loops, cliques or some other complex structures.

Situation with a loop when paper *A* cites paper *B* and paper *B* cites *A* is theoretically possible, for example if authors exchange their deliverables and cite not yet published but already accepted for publication papers. However, according to Glänzel [9] traditional scientometrics does not consider such citations as the valid ones.

2.2 Focused Surfer

The Random Surfer model is the basis of PR algorithm. Page Rank of the certain node is proportional to the probability to reach this node by randomly riding the graph. At each step rider randomly chooses the link to follow. Focused Surfer decides which path is more preferable for him. Formula (4) expresses this mathematically:

$$P_i = (1-d) \cdot \sum_{i \neq j}^{j \in D} P_j \cdot s(j \mid i) + \frac{d}{N}, \quad (4)$$

where *s(j|i)* is the probability to follow the reference *i* being at the place *j*. *s* is a function that may be arbitrary. We propose to use the simplest variant of it, which we show in formula (5):

$$s(j \mid i) = \frac{C(i)}{\sum_{k \in D} C(k)}, \tag{5}$$

where *C(m)* is paper *m* citations count, and *D* is the set of all references in paper *C(j)*. This means that more cited nodes have advantage and they are more visible and attractive for further citation.

3. Evaluation and experimental methodology

In our evaluation, we explore 266788 papers published in ACM conferences or journals starting from 1950 and till 2007 with the majority of papers around 2002-2005. This dataset may be completely matched to ACM portal[2] and was crawled by the Citeseer[3] digital library.

3.1 Plotting the difference

We introduce here our proposed experimental methodology. The obvious approach to exploring the effect of using PR vs citation count (CC) in evaluating papers is to plot these values for the different papers. The density of points (points cloud) that have a high CC and low PR (or vice versa) would provide an indication of how often these measures can give different quality indication for a paper. However, this leads to charts difficult to read in many ways: First, points overlap because many papers have the same CC, or the same PR, or both. Second, it is hard to get a qualitative indication of what is "high" and "low" for CC or PR. This is why we divide CC and PR axis in bands.

Ideally we would have to split the axes into 10 (or 100) bands. We put in the first band the top 10% (top 1%) of the papers based on the metric, to give qualitative indications so that the presence of many papers in the corners of the chart would denote a high divergence. However, the overlap problem would remain, and it would distort the charts in a significant way since the measures are discrete. For example, the number of papers with 0 citations is well above 10%. If we neglect this issue and still divide in bands of equal size (number of papers), papers with the same measure would end up in different bands.

Finally, the approach we took (Fig. 1, Fig. 2) is to divide the X-axis in bands where each band corresponds to a different - discrete - citation count. With this separation we built 290 different bands, since there are 290 different values for CC (even if there are papers with much higher CC, there are only 290 different CC values in the set). For the Y-axis we leverage mirrored banding, i.e., the Y-axis is divided into as many bands as the X-axis, also in growing values of PR. Each Y band contains the same number of papers as X. In other words, the vertical rectangle corresponding to band *i* in the X axis contains the same number of papers q_i as the horizontal rectangle

[2] http://portal.acm.org/
[3] http://citeseer.ist.psu.edu/

corresponding to band *i* of the Y-axis. We call a point in this chart as a square, and each square can contain zero, one, or many papers (not ranks, because the zone number represents the actual PR or CC).

The reasoning behind the use of mirrored banding is that this chart emphasizes divergence as distance from the diagonal. At an extreme, plotting a metric against itself with mirrored banding would only put papers in the diagonal. Since the overlap in PR values is minimal (there are thousands of different values of PR and very few papers with the same PR values, most of which having very low CC and very low PR, and hence uninteresting), it does not affect in any qualitatively meaningful way the banding of the Y-axis. To realize what are the real value of PR and CC is behind of each zone please take a look at the Table 1.

Table 1. The mapping between real CC and PR and the band number.

Number of band both for CC and PR	CC	PR
50	50	6.23
100	100	14.74
150	151	26.57
200	213	38.82
250	326	58.86
280	632	113.09
290	1736	224.12

3.2 Evaluation

The described analysis and visualization methodology gives the overall picture for all 266788 papers on one chart (Fig. 1). The points are strongly biased around the main diagonal. This biasing shows the *diversity*, or difference between PR and CC. There are some papers with extremely low citation count but very significant Page Rank, or "scientific gems" following Chen *et al.* [3]. They are the papers cited by several heavily cited papers. Being cited by just *one* extremely high ranked paper may be enough to improve PageRank drastically. Fig. 3 represents a piece of full citation graph, where there is a real paper with PR>>CC and just 14 citations from other papers in the graph.

In contrast to "scientific gems" there are some other papers below the main diagonal, located in the bottom-left part of Fig. 1 and Fig. 2, when CC band is greater than 50. Papers in that region have significantly high CC and small PageRank. This is caused by "outgoing links effect". To understand the nature of this effect let us see the formula (1). Denominator *S* in (1) represents the probability to follow the link, and being a big number it reduces the Page Rank of a paper. This denominator *S* is the corner-stone of Random Surfer model, and it reflects the fact that all references are completely equal from probabilistic point of view. Thus if a paper is cited many times by papers with large quantity of *outgoing links* (papers with long "references" section) it may have much lower PageRank than the other papers with the same citation count. The example of such a paper is plotted in the middle of Fig. 4, this paper has 55 citations and more than 100 times lower PR than "scientific gem" plotted in Fig. 3.

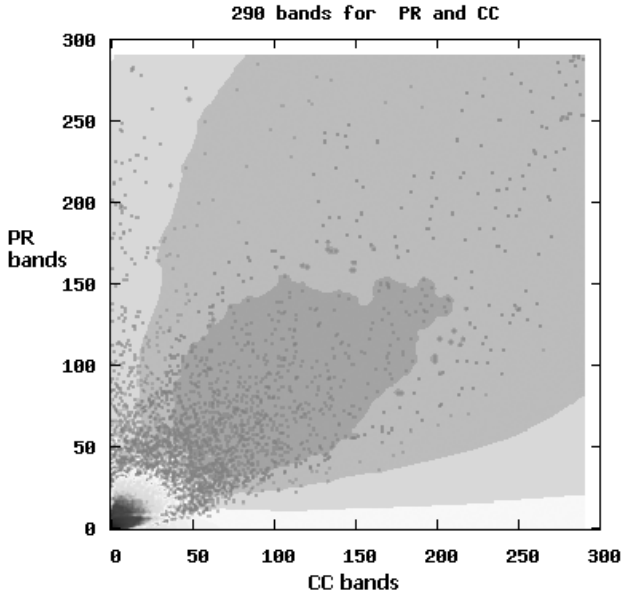

Fig. 1. Diversity of Page Rank (PR) and Citation Count (CC). White and black points in the bottom-left corner does not mean absence of papers. This is a grayscale of colored map, where the major quantity of papers has small number of CC, and since lie exactly in the bottom-left corner and it is nearly the same for the both plots. The plot is mirror-like banded.

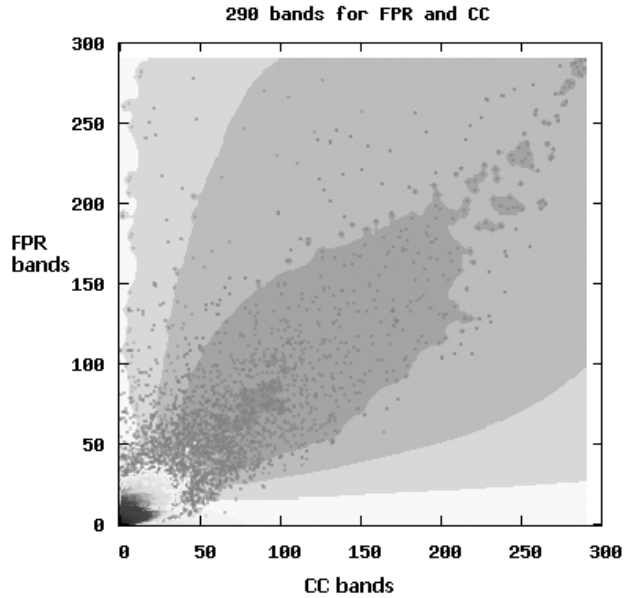

Fig. 2. Diversity of Focused Page Rank (FPR) and Citation Count (CC). Bottom-left corner distribution explained on the **Fig. 1** description. Again the plot is mirror-like banded.

Fig. 2 illustrates the Focused Surfer model and FPR algorithm instead of PR. Focused Surfer model gives better chances to more cited papers, at the same time stealing the part of the weight from their poorly cited neighbors. This idea leads us to the conclusion that in general total FPR rank remains the same as PR, it just gets *re-distributed*. This idea is supported by computation of average FPR and PR which are nearly the same: $<FPR>=0.603$ and $<PR>=0.602$.

Now let us observe effects present on Fig. 2. The points are located closer to the main diagonal[4] (comparing with Fig. 1) and there is significantly less papers with big CC and small PR (reducing of the effect of outbound links). On the other hand we see that "gems"-effect is still noticeable.

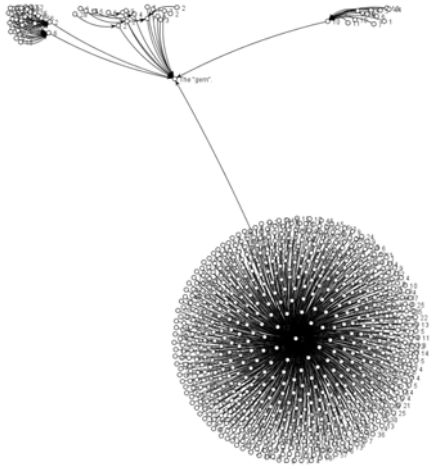

Fig. 3 "Scientific gem" in the center. Cited by heavily cited paper (in the bottom).

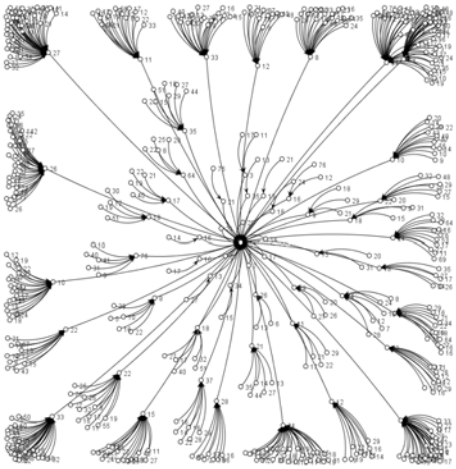

Fig. 4. The opposite to "scientific gem" paper (in the middle).

[4] If all the points lie exactly on the main diagonal we would have 100% match of CC and PR.

This means that FPR tends to reduce "outgoing links" effect and tends to make FPR closer to the CC. Or in other words it tends to shift the points in Fig. 2 towards the main diagonal. This effect of *shifting* the points towards the main diagonal may be numerically evaluated. We compute the difference $\Delta_{gems}=PR-FPR$ for each *square* in the "gems zone", where CC band<10 and PR band >10. Then we do the same for the opposite "popular papers" zone where CC band>10 and PR band<10. It would be $\Delta_{popular\ papers}=PR-FPR$. We notice that $\Delta_{gems}=3\ \Delta_{popular\ papers}$, which means that focusing eliminates "popular papers" 3 times greater than "gems". So Focused Page Rank tends to keep "gems" while correcting "popular papers" ranks.

The last plot in Fig. 5 shows the top 100 papers with the biggest CC. There are 3 curves there: PR, CC and FPR. It is clear from Fig. 5 that FPR is a tradeoff between PR and CC in highly cited region.

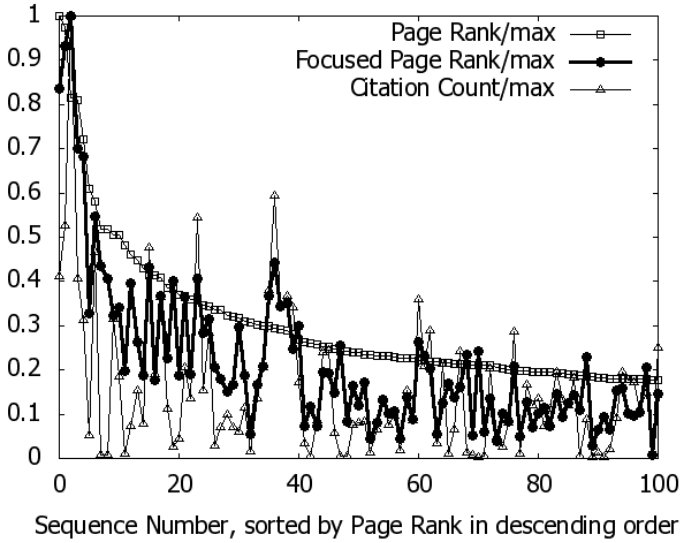

Fig. 5. Top 100 papers with the highest CC. Bold line is the Focused Page Rank. All ranks are normalized by their maximum value, and thus comparable.

4. Conclusion

Focused Page Rank has been proposed for the problem of scientific citing. Our major strong points are:

1. *It is the tradeoff between Page Rank and Citation Count*. So it may serve as an agreement between the followers of pure citation count and Page Rank followers.
2. The proposed Focused Page Rank suffers less from the effect of outbound links. Therefore, it is capable to better capture one of the fundamental principles of Scientometrics, first time formulated by de Solla Price in 1976 [11]:

   "Success seems to breed success. A paper which has been cited many times is more likely to be cited again than one which has been little

cited. An author of many papers is more likely to publish again than one who has been less prolific. A journal which has been frequently consulted for some purpose is more likely to be turned to again than one of previously infrequent use".

3. It captures the power of Page Rank, where not only the quantity of citations, but also the quality of ones counts.

5. Acknowledgements

Author would like to thank Prof. Fabio Casati, Andrei Yandratsau and Alexander Autayeu for useful discussions, and Prof. Lee Giles for providing high quality dataset.

6. References

1. Page L., Brin S.: The anatomy of a large-scale hypertextual web search engine. Proceedings of the Seventh International Web Conference, pp. 107 - 117 (1998)
2. Diligenti M., Gori M., Maggini M.: Web Page Scoring Systems for Horizontal and Vertical Search. In: WWW2002, pp. 84-89. ACM Press, New York (2002)
3. Chen P., Xie H., Maslov S., Redner S.: Finding scientific gems with Google's PageRank algorithm. Journal of Informetrics, v. 1, n. 1, pp. 8-15. (2007)
4. Haveliwala T.: Efficient Computation of PageRank. Techical report, http://dbpubs.stanford.edu/pub/1999-31, pp. 84-89 (1999)
5. Langville A. N., Meyer C. D.: Deeper Inside PageRank. J. Internet Mathematics, v. 15, n. 5, pp. 335-380, (2004)
6. Kamvar S., Haveliwala T., Manning C., Golub G.: Extrapolation Methods for Accelerating PageRank Computations. In: WWW2003, ACM 1581136803/03/0005 (2003)
7. Abou-Assaleh T., Das T., Weizheng G., Yingbo M., O'Brien P., Zhen Z.: A Link-Based Ranking Scheme for Focused Search. In: WWW2007, ACM Press. 2007
8. Fuyong Y., Chunxia Y., Jian L., WImprovement of PageRank for Focused Crawler. In : Eighth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, pp. 797-802 (2007)
9. Glänzel, W. Bibliometrics as a research field, A course on theory and application of bibliometric indicators, Magyar Tudományos Akadémia, Course Handouts http://www.norslis.net/2004/Bib_Module_KUL.pdf , (2003)
10. Sun Y., and Giles C. L.: Popularity Weighted Ranking for Academic Digital Libraries, 2007
11. de Solla Price D. J.: Little Science - Big Science. , Columbia Univ. Press, New York, (1963)
12. Bollen J., Van de Sompel H., Balakireva L., Chute R.: A ranking and exploration service based on large-scale usage data. In: JCDL, ACM/IEEE, poster (2008)
13. Sobek M. The effect of outbound links, Internet paper. http://pr.efactory.de/e-outbound-links.shtml, (2003)