8th International Conference on Applied Informatics
Eger, Hungary, January 27–30, 2010.

PageRank Based Network Algorithms for Weighted Graphs with Applications to Wine Tasting and Scientometrics

Tibor Csendes[a] and Elvira Antal[a]

[a]Institute of Informatics, University of Szeged, Hungary
e-mail: csendes@inf.u-szeged.hu

Abstract

The PageRank algorithm was originally designed to determine the importance of home pages, and is a successful part of the Google search engine. A generalization of it for weighted graphs have been considered, when the edges of the graphs are appropriately weighted to provide additional information when the connections have different meaning, importance or quality.

The algorithm was applied for rankings obtained in wine tasting to determine the quality or expertise of wine taster – providing a novel, objective methodology for extracting merit. Results are reported for the case of the Szeged Wine Fest competition data.

A similar algorithm will be discussed with applications in scientometrics to give a new measure of the quality of scientific publications – based on the citation structure. Results are reported for the scientometric qualification of the publications of Jenő Egerváry.

Keywords: PageRank algorithm, scientometrics, wine tasting

MSC: 05C85, 90C35

1. Introduction

The PageRank algorithm was developed by the founders of Google to calculate a good approximation of the importance of web pages. This measure was then used to rank the found pages for search results. The PageRank algorithm works on a directed graph. Each node has the same value at the beginning. The procedure will simulate the behaviour of an average web surfer: starting from a random page, it jumps to one of those pages which are linked to the given one. After every sixth passage along links, the surfer jumps randomly to another page. Pages that are often visited by such random surfers are regarded as high ranking [4, 9].

The formal algorithm sets the values of nodes to one at the beginning. Then, in each iteration step, the value of each node will be determined by the value of those

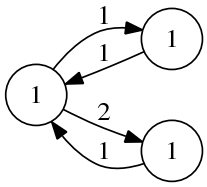

Figure 1: A simple example of a weighted graph.

	iteration number								
node	1	2	3	4	5	6	7	8	9
1	1.00	1.83	1.14	1.71	1.24	1.64	1.30	1.58	1.35
2	1.00	0.44	0.68	0.48	0.64	0.51	0.62	0.53	0.61
3	1.00	0.72	1.19	0.80	1.12	0.85	1.08	0.89	1.05

Table 1: The result of a few iterations of the PageRank algorithm on the weighted graph given on Figure 1.

nodes from where a link, a directed edge of the graph leads to the given node. The random jumps are also simulated. The iteration expression is

$$\mathrm{PageRank}(i) = \frac{1-d}{N} + d \sum_{j \in M(i)} \frac{\mathrm{PageRank}(j)}{L(j)},$$

where $1-d$ is a damping factor indicating the probability of random jumps, N is the number of nodes considered, $M(i)$ is the set of those nodes that have links to the node i, and $L(j)$ is the number of outgoing links from node j. After a few iterations, the values of the nodes usually converge to values that reflect the frequency with which they are visited by a random surfer.

We can generalize this approach by allowing weights for the links (or edges) [8] reflecting how important these are. In the present paper we consider the generalization of PageRank to weighted graphs. The main difference is that the outgoing links of a node should not have necessarily the same strength. In our applications, this new feature will be crucial. In the case of wine tasting, the coherence of the points given for a particular wine will make the given link stringer. Also in scientometric analysis [7], there is an obvious additive meaning of a definitely more important citation that can be easily modeled by a link with larger weight. A link with a weight 2 means basically nothing else as if this edge would be doubled, see the simple example on Figure 1.

The results of a few iteration of the PageRank algorithm on the weighted graph given by Figure 1 confirm our expectation that node 3 (the one where the link with weight 2 leads) has more merit value than node 2.

2. Application to wine tasting

Wine tasting is a mostly personal, subjective procedure to determine the quality of wines. Several wines are usually evaluated in one sitting in an anonymous way, called blind tasting (the tasters do not know which wine is the next one). The wines obtain points from the tasters and the ranking of the committee of a few taster is possibly repeated if a close decision is met. It is open how to determine the quality of the tasters, no objective measure or procedure is known to us.

Let us see now how the weighted PageRank algorithm can serve wine tasting. We start from the same merit value for each wine taster, and in each iteration cycle we consider the points given for the next tasted wine. The basis information to set the weight of the link between two wine taster is the difference of their evaluation value. We repeat the procedure until the merit values stabilize.

It is easy to recognize that the underlying assumption is that good wine tasters have the capability of a sure, precise evaluation. In other words, those tasters that give mostly coherent ranking, similar points to the same wines, must be the good ones. While those who have usually substantially different rankings compared to other tasters – must be less qualified. It is not at all obvious whether this assumption really holds. Only results obtained by applying this assumption can confirm whether it was a good idea to base the new procedure on this easy to obtain information.

We define the mathematical model as follows. The directed graph to be studied is the complete directed graph between the wine tasters as nodes. Let us now define the weights of the edges. For wine k the link between the wine tasters i and j obtains the weight

$$\frac{1}{N(|p(i,k)-p(j,k)|+1)},$$

where $p(i,k)$ is the number of points given by wine taster i for wine k, and N is the number of wines tasted. Closely equal evaluation point values will produce a large weight between the related tasters.

To test our algorithm and model, we have selected the data of a recent wine tasting event. At the Szeged Wine Fest 2009, 104 wines were blind tested by 4 groups of tasters (5 persons in each team). The testing teams obtained 33-34 different wines. Each wine obtained some points in the international 100 point evaluation system. The normalized wine taster rankings were obtained by our algorithm for team 1 as 0.954, 1.000, 0.889, 0.894, and 0.884. Although the evaluation was made anonymously, the first figure given above belongs to a wine producer of the year. Although a very good wine producer is not necessarily a good wine taster, still a kind of positive conclusion can be drawn on this basis: our methodology produces realistic looking results for the quality of wine tasters.

The detailed results for the Szeged Wine Fest data are summarized on Table 2. The calculated numbers can be interpreted as normed merit values: the larger the better.

taster	team 1	2	3	4
1	0.954	0.955	1.000	1.000
2	1.000	0.838	0.879	0.995
3	0.889	0.905	0.854	0.929
4	0.894	0.855	0.891	0.952
5	0.884	1.000	0.934	0.978

Table 2: The detailed results on the 2009 Szeged Wine Fest data.

taster	team 1	2	3	4
1	1.0000	0.9962	1.0000	0.9948
2	0.9963	0.9812	0.9888	0.9989
3	0.9922	0.9863	0.9842	0.9932
4	0.9917	0.9907	0.9854	0.9943
5	0.9909	1.0000	0.9882	1.0000

Table 3: The detailed results on the 2009 Szeged Wine Fest data.

We have repeated the computational study with a simplified algorithm. This has not considered the evaluation point differences one by one, but in stead, the summarized differences between the evaluation points were calculated, and the iteration was made using this single weight system (in contrast to the previous approach, where the results off wines were handled separately). Results for the Szeged Wine Fest data based on summed differences are given in Table 3. The less detailed information produced less differentiated results. The results of the two evaluations differ slightly. It is reasonable, and we can draw the conclusion that both evaluations methods can be useful – after a proper practical comparison with common sense rankings.

3. Application to scientometrics

Scientometrics aims to measure the amount and value of scientific work done as reflected in scientific publications. The main tool used is citation analysis. Usually the number of citations for a paper is regarded to be proportional to the value of the publication and to the novelty involved. The impact factor of a journal is a measure of importance again: it is the expected number of citations a paper in the given periodical will obtain within two years.

These indicators should usually handled by care, they are much dependent on the subject area, and are regarded as reliable only for larger sets of data (better just for institutes than for individuals). Still, much criticism can be read on this

methodology.

The simple application of the PageRank algorithm for the directed graph of scientific publications is an accepted scientific merit evaluation methodology (when the links are defined by citations). In addition to that, weighted PageRank allows us to handle citations of different importance.

Obviously, when a citation is just one of many in a row, giving only possible connection points, then it has less relevance to the presented new result compared to those references that are termed to have a direct connection to the present paper, or even forming an important basis for it. Such weighting can be established on an open access, but recorded and moderated peer review basis (such as that used for Wikipedia), or can even be composed by artificial intelligence algorithms. These can produce good estimations of such weighting.

To see the capabilities of our new approach, we have tested the method on the famous paper of Jenő Egerváry [2]. As it is widely known, Harold Kuhn developed an algorithm for the solution of the assignment problem [6], and called it the Hungarian method acknowledging the contributions of Jenő Egerváry and Dénes König [2, 5], who wrote important papers in Hungarian and in German that contained important theoretical basis for the later algorithm. What is more, Kuhn even learned the Hungarian language to a modest level, and translated the paper of Egerváry using a dictionary.

According to classic scientometric evaluation, the paper of Egerváry received just a few citation, while some of the citing papers much more. For example, the ISI Web of Knowledge gives 38 citations for the paper of Egerváry, while 726 for that of H.W. Kuhn and 215 for the mentioned paper of D. König. As we shall see, the new PageRank based method showed much larger reputation for the publication of Jenő Egerváry.

We have completed two experiments. In both cases we collected scientometric data from the ISI Web of Knowledge database. In the first setting we have selected the seminal paper of Egerváry [2], those in the database which have cited it, and also those which were cited be the paper of Egerváry. Then we have established the citation relations among the papers. The resulting directed graph can be seen on Figure 2. Here node number 1 in the center represents the paper of Egerváry, and node number 43 (upper left to the previous) that of Kuhn. Then we have set the merit values of the papers to the number of citations available in the same database, and fixed these values by rewriting the same numbers after each iteration – with the exception of the node related to the paper of Egerváry.

The idea behind this procedure was the observation that the restriction of the complete graph to a subset of it will produce the same result as the original – assuming that the boundary of the subgraph has the converged merit values of the full graph. Obviously, it cannot be ensured without having run the algorithm on the full graph, still it seems to be an acceptable approximation to use the traditional citation numbers in stead. The applicability of this assumption will be justified again by reasonable results obtained for the subgraph.

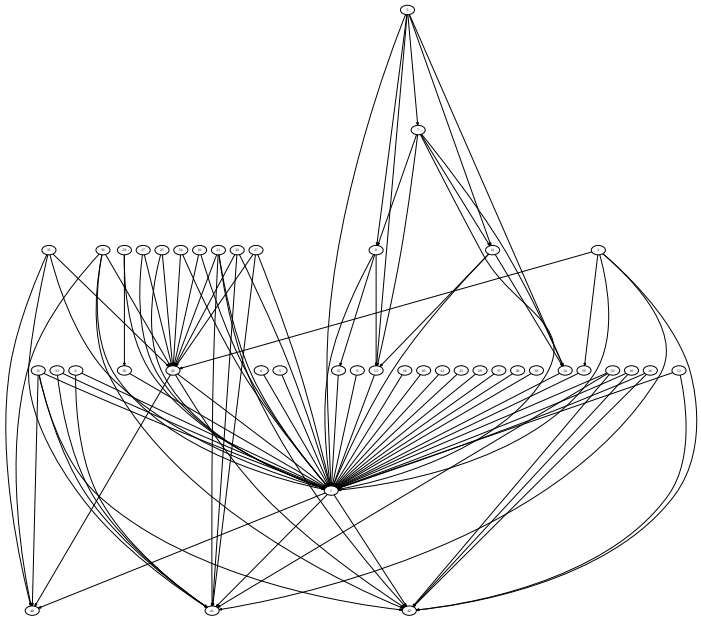

Figure 2: The subgraph studied in connection to the seminal Egervary paper.

The executed Matlab procedure basically just applied the power method to determine the eigenvalues of the adjacency matrix of the graph:

```
a=load('31_Matrixok_Egervary_1szint.csv');
N=max(max(a));
A=sparse(a(:,1),a(:,2),ones(size(a,1),1),N,N);
A=A';
A=colnorm(A); clear a;
clear N;
w=load('31_M_E_1szNR.csv')

[y0, y, nu, iter]=powmeth(A,w,100,0.00001)
```

The normed citation counts for the related papers, and their converged values after running 18 iterations of the PageRank algorithm are given in Table 4. This result can be interpreted as the scientometric merit value of the paper by Egerváry has been increased by 7.5 times.

We have repeated the same procedure for the larger subset of the citation graph, that contained also those papers that have cited the nodes of the earlier test. Then, again we have explored all the interconnections among the considered papers, and fixed the boundary of the subgraph, while the inner nodes could change as the

iteration	Egerváry	...	Kuhn
0	0.0451	...	0.8616
18	0.3383	...	0.8616

Table 4: The normed citation counts for two papers of the citation subgraph around the paper of Egerváry.

iteration	Bellman	...	Egerváry	...	Ford-Fulkerson	...	Kuhn
0	0.0133	...	0.0282	...	0.0289	....	0.5384
10	0.3770	...	0.2513	...	0.2343	...	0.6805

Table 5: The normed citation counts for four papers of the citation extended subgraph around the paper of Egerváry.

PageRank algorithm produced. We call this subgraph second level citation graph of the paper of Egerváry. The subgraph contained over 1000 nodes, we do not give a figure on that (it would be too complex and would deliver juts minimal additional information). The result is summarized in Table 5. Obviously, these results should be considered as more realistic, since a more detailed structure of the citation graph has been investigated, and also we have used more data in this calculation compared to the first study.

This second set of result data indicate again the improvement of the scientific value of the paper of Egerváry, this time almost 9 times. Although its merit value remained below that of the paper of Kuhn, still it become more realistic, and comparable to common sense. It is interesting to see, that also the scientific merit value of the paper of Kuhn increased (to a less extent) after the iterations. Just to additional papers could improve it merit value in this setting: that of Ford and Fulkerson [3] (from 0.0289 to 0.2343) and the one by Bellman [1] (from 0.0133 even to 0.3770).

Please note that although our computational procedure would allow, now we have not considered different weights for the citations of a paper. This task remains for future research.

As a conclusion, we can summarize our experiments that the suggested weighted PageRank algorithm produced promising results both on the wine tasting and on the scientometric data, and further investigations can clarify its future role and the suitable algorithm details to enable efficient and informative application.

Acknowledgements. This work was supported by the grant TÁMOP-4.2.2/08/1/2008-0008 of the Hungarian National Development Agency. The authors are grateful to Márk Jelasity for his advices and to Melinda Braun for the tedious work of typing the wine tasting data.

References

[1] Bellman, R., J. of the Society for Industrial and Applied Mathematics 4(1956) 168–205

[2] Egerváry, J., Mátrixok kombinatórius tulajdonságairól (On the combinatorial properties of matrices, in Hungarian), Matematikai és Fizikai Lapok 38(1931) 16–28

[3] Ford, L.R., Fulkerson, D.R., Solving the transportation problem, Management Science 3(1956) 24–32

[4] Kleinberg, J.M., Authoritative Sources in a Hyperlinked Environment. J. of the ACM 46(1999) 604–632

[5] König, D., Über Graphen und ihre Anwendung auf Determinantentheorie und Mengenlehre. Mathematische Annalen 77(1916) 453–465

[6] Kuhn, H.W., The Hungarian Method for the assignment problem, Naval Research Logistics Quarterly 2(1955) 83–97

[7] Nan Ma, Jiancheng Guan, Yi Zhao, Bringing PageRank to the citation analysis. Information Processing and Management: an Int. J. 44(2008) 800–810

[8] Nemirovskya, D., Avrachenkov, K.,: Weighted PageRank: cluster-related weights. In Proc. of the 2008 Text REtrieval Conference (TREC 2008), Gaithersburg, MD, 2008

[9] Parreira, J.X., Donato, D., Michel, S., Weikum, G.,: Efficient and Decentralized PageRank Approximation in a PeertoPeer Web Search Network. Proc. of the 32nd Int. Conf. on Very Large Data Bases, Seul, Korea, 2006, 415–426

Tibor Csendes
H-6720 Szeged, Árpád tér 2, Hungary