*Annales Mathematicae et Informaticae*
*44 (2015) pp. 131–140*
http://ami.ektf.hu

# A local PageRank algorithm for evaluating the importance of scientific articles[*]

**András London[†‡], Tamás Németh**
**András Pluhár, Tibor Csendes**

Institute of Informatics, University of Szeged, Hungary
london@inf.u-szeged.hu

*Submitted July 15, 2014 — Accepted March 5, 2015*

**Abstract**

We define a modified PageRank algorithm and the $PR$-score to measure the influence of a single article by using its local co-citation network. We also calculate the reaching probability and $RP$-score of a paper starting at an arbitrary article of its co-citation network for the same purpose. We highlight the advantages of our methods by applying them on the celebrated paper of Jenő Egerváry that is underrated by the standard indices.

*Keywords:* Scientometric, PageRank, Ranking algorithms, Co-citation networks

## 1. Introduction

The relevance of scientometrics – aiming at measuring the productivity and quality of scientific research – has long been widely discussed in the academic domain. Among the most popular measures used are scientific citation indices due to their easy accessibility. Several of these indices have been introduced such as $h$-index (or Hirsch-index) proposed by Hirsch [14], the $g$-index proposed by Egghe [11],

[*] This work was partially supported by the European Union and the European Social Fund through project FuturICT.hu (grant no.: TAMOP-4.2.2.C-11/1/KONV-2012-0013).

[†] The first author was supported by the *European Union* and the *State of Hungary, co-financed by the European Social Fund* in the framework of TAMOP-4.2.4.A/2-11-1-2012-0001 'National Excellence Program'.

[‡] Corresponding author: london@inf.u-szeged.hu

the $w$-index and maximum-index both proposed by Woeginger [27]. All of these indices are based on the citation records of the researchers. These indices have been extensively criticized since they are much dependent on the scientific field (e.g. number of researchers and available journals, popularity of the area, gender ratio, etc., see e.g. [1, 19, 26]). Another drawback is that the number of citations does not give a clear picture on the influence and quality of a single paper.

Several studies have been addressed this problem using network approach. Co-citation networks, in which nodes represent single articles and a directed edge represents a citation from a citing article to a cited article, describes the relation between citations of different papers have been widely studied previously[6, 15, 20]. Chen et al. [7] applied the PageRank algorithm [5] (developed by the founders of Google) for co-citation networks, later Raddichi et al. [23] defined an iterative ranking method analogous to different ranking algorithms such as PageRank, CiteRank [25] and HITS [16] in order to evaluate the influence of single articles by using co-authorship networks (where nodes represent publications and weighted edges represent the number of common authors of them). Several modifications and variants of the network models have been designed in the context of scientometrics (see e.g. [12, 21, 24, 28]).

More recently, the Eigenfactor Score and the Article Influence Score [4] have been developed to estimate the relative influence of single articles based on citation networks as well. Furthermore the underlying algorithms can also be applied to journals, authors, and institutions.

Following the network approach, our main goal is to measure the influence of a single article regardless of the specialties of the field. Based on the previous results of Csendes and Antal [9] and by applying the experimental results of Chen et al. [8] that is later mathematically proved to be efficiently applicable for many classes of graphs by Bar-Yossef and Mashiach [2], we use a local PageRank estimating method for this purpose. It is important to note that we do not want to attempt to determine the scientific value of the articles (which will be probably judged in the future).

This article is organized as follows: in Section 2 we give a brief mathematical overview of the PageRank algorithm. In Section 3, we describe how a local PageRank method can be applied to determine the scientific influence of a research paper. Finally in Section 4 we compute the local PageRank values of the articles in the co-citation graph of the famous paper of Jenő Egerváry [10] and highlight the main advantages of our approach from the scientometric point of view.

## 2. Methods

In this section, we give a short mathematical overview of the PageRank algorithm. We describe the main notions, definitions, and theoretical results of a local PageRank method that we used. We omit the proofs of the theorems that can be found in [2].

## 2.1. Overview of the PageRank method

The PageRank algorithm was originally designed to provide a good approximation of the importance of web pages. Since it works on directed graphs, it is a natural idea to use the PageRank method for ranking the articles in co-citation graphs.

Let $G = (V, E)$ be a directed graph of $N$ nodes. Let $d^-(i)$ $(i = 1, 2, \ldots, N)$ be the number of outgoing edges from a node $i$ and $N^+(i) = \{j \in V : j \to i \text{ exists}\}$, i.e. the set of nodes having an edge to node $i$. PageRank of a node $i \in V$ is defined then by the following recursion formula [5]:

$$PR(i) = \frac{\lambda}{N} + (1-\lambda) \sum_{j \in N^+(i)} \frac{PR(j)}{d^-(j)}, \tag{2.1}$$

where $\lambda \in [0, 1]$ is a free parameter (usually set between 0.1 and 0.2).

The PageRank formula defined by equation (2.1) can be written in vector equation form, and then the PageRank vector **PR** is defined as

$$\mathbf{PR} = \frac{\lambda}{N}[I - (1-\lambda)AD^{-1}]^{-1}\mathbb{1}, \tag{2.2}$$

where $A$ is the adjacency matrix of $G$, $D$ is a diagonal matrix such that $D_{ii} = \sum_{\ell=1}^{N} A_{i\ell}$ and $D_{ij} = 0$, if $i \neq j$, $I$ is the $N \times N$ identity matrix and finally $\mathbb{1}$ is the $N$-dimensional vector having each component equals to 1.

Assuming that $\mathbb{1}\mathbf{PR} = 1$, Eq. (2.2) implies, that

$$\mathbf{PR} = [\frac{\lambda}{N}\mathbb{1}\mathbb{1}^T - (1-\lambda)AD^{-1}]\mathbf{PR}, \tag{2.3}$$

which shows, that **PR** is the eigenvector of the matrix $\frac{\lambda}{N}\mathbb{1}\mathbb{1}^T - (1-\lambda)AD^{-1}$ due to the fact that an eigenvalue equals to 1, which is the largest eigenvalue of this matrix by a consequence of the Frobenius-Perron theorem for row-stochastic matrices [22].

More intuitively, let us consider a random walk on the nodes of the graph. Starting from a node $i$, a random surfer selects one of the node's outgoing edges randomly with uniform distribution, moves to the end node $j$ of that edge, and repeat this process from $j$, etc. The parameter $\lambda$ can be understood as a "damping" factor which guarantees that the random walk restarts in a random node of the graph, chosen uniformly random, almost surely in every $1/\lambda$-th step. This can guarantee, that the process would not stop by reaching a node with an out-degree zero. If the surfer reaches a node, the number of visits of that node increases by one. The damping factor ensures that each node receives a contribution $\lambda/N$ at each step. Thus, the PageRank of a node $i$ can be considered as the long-term fraction of time spent in node $i$ during the random walk. The steady-state of the random walk is given by the solution of Eq. (2.3).

## 2.2. Local PageRank approximation

Although in many applications PageRank scores are needed to be computed for all nodes of the graph, there are situations in which one is interested in computing

PageRank scores only for a small subset of the nodes. Chen et al. [8] developed an algorithm to approximate the PageRank score of a target node of the graph with high precision. Their algorithm crawls backwards a small subgraph around the target node(s) and applies various heuristics to calculate the PageRank scores of the nodes at the boundary of this subgraph and then computes the PageRank of the target node(s) by using only the crawled subgraph. By using simulations, they showed that this algorithm gives a good approximation on average. On the other hand, they also pointed out that high in-degree nodes could make the algorithm very expensive and incorrect.

From now in this section, we use the same notions as in [2]. An algorithm is said to be an *$\epsilon$-approximation* of the PageRank, if for a graph $G = (V, E)$, a target node $i \in V$ and a given error parameter $\epsilon > 0$, the algorithm outputs a value $PR'(i)$ satisfying

$$(1-\epsilon)PR_G(i) \leq PR'(i) \leq (1+\epsilon)PR_G(i). \tag{2.4}$$

For a directed path $p = (k_1, \ldots, k_t)$ from node $k_1$ to $k_t$, let $w(p) = \prod_{i=1}^{t-1} \frac{1}{d^-(k_i)}$, that is the reaching probability of $k_t$ from $k_1$ in a given path, where the transition probabilities are proportional to the number of outgoing edges. Let $p_t(i,j)$ be the set of all directed path of length $t$ from $i$ to $j$. Then, the *influence* of node $i$ on the PageRank of node $j$ at radius $t$ is defined as

$$I_t(i,j) = \sum_{p \in p_t(i,j)} w(p), \tag{2.5}$$

and thus, the total influence of $i$ on $j$ is

$$I(i,j) = \sum_{t=0}^{\infty} I_t(i,j). \tag{2.6}$$

By using the definition of influence, PageRank of node $j$ at radius $r$ can be defined as

$$PR_G^r(j) = \frac{\lambda}{N} \sum_{t=0}^{r} \sum_{i \in V(G)} (1-\lambda)^t I_t(i,j). \tag{2.7}$$

It can be proved that for every node $j \in G$, $PR_G(j) = \lim_{r \to \infty} PR_G^r(j)$ holds (the proof can be found e.g. in [2]). The interesting question is that how small the radius $r$ can be such that the PageRank approximation would even be appropriate.

In [2] it was proved, that the hardness and inappropriate nature of local approximation of PageRank on certain graphs (constructed examples) is caused by two factors: the existence of high in-degree nodes and the slow convergence of PageRank iteration algorithm. We shall see, that in our case (and in most of the co-citation graphs in scientometrics) these properties does not hold.

It was also shown, that the several variants of the approximation algorithms proposed by Chen et al. are still efficient on graphs having bounded in-degrees and admitting fast PageRank convergence.

Let us be given a $G = (V, E)$ graph, node $j \in V$ and the approximation parameter $\epsilon$. The *point-wise influence mixing time* of $j$ is defined as

$$T_G^{\epsilon}(j) = \min\{r \geq 0 : \frac{PR_G(j) - PR_G^r(j)}{PR_G(j)} < \epsilon\}. \tag{2.8}$$

The algorithm we use computes $PR_G^r(j)$ for a given node $j$ (see in Section 4) and it follows from the definitions that it runs with $r = T_G^{\epsilon}(j)$ and gives an $\epsilon$-approximation of $PR$. To complete the description of the theoretical background, we should see the upper bound on $T_G^{\epsilon}(j)$ (or radius $r$).

For graph $G = (V, E)$ with $j \in G$ and $r \geq 0$ the *crawl size* at radius $r$ is defined as

$$C_G^r(u) = \#\{i \in G : \exists p_t(i, j) \text{ with } t \leq r\}. \tag{2.9}$$

It is immediate from the definition, that if the local PageRank algorithm runs for $r$ iteration, its cost is $C_G^r(u)$. A trivial upper bound for the crawl size is that $C_G^r(u) < d^r$, where $d$ is the maximum in-degree of $G$.

Finally, it was also proved that for any $G$ directed graph, node $j \in G$ and $\epsilon > 0$ it holds that a radius $r = \mathcal{O}(\log(1/PR_G(u)))$ is always sufficient (while in practice a much lower radius could be enough).

## 2.3. Reaching Probabilities

A possible simplification of the PageRank method is to consider only the reaching probabilities of the nodes in the network. We would like to know the probability of reaching a node $j$ starting from an arbitrary chosen node $i$ of the network. The reaching probability, $RP$ of node $j$ can be defined as

$$RP(j) = \sum_{i \in N^+(j)} p_{ij} RP(i), \tag{2.10}$$

where $p_{ij}$ is the reaching probability of node $j$ from a neighbor node $i$. It is natural to assume, that each possible selection of a neighbor of node $i$ has equal probability, thus we can write $p_{ij} = 1/d^-(i)$ in Eq. (2.10). By this choice, Eq. (2.10) is the PageRank equation without the damping factor. However, in contrast to the calculation of PageRank, we do not want to evaluate the vector $RP$ in the steady-state. Instead, we only determine the reaching probability of a given node $j$, which can be calculated as

$$RP(j) = \frac{1}{N} \sum_{i \in V} I(i, j), \tag{2.11}$$

where $I(i, j)$ is as defined in (2.6). In the point of view of published articles, $RP$ can be interpreted as the probability of a given article can be found by someone (e.g. a scientist), who starts the search at any article and goes to another randomly chosen article cited by the current one.

## 3. Application to scientometrics

In the last decade, co-citation networks have been investigated aiming to measure the importance of a scientific article. A co-citation network is defined as a directed graph $G = (V, E)$ of $N$ nodes, where each node $i \in V$ refers to an article and there is a directed edge $i \to j \in E$ from node $i$ to node $j$ if article $j$ is cited in article $i$. Our method, that aims to measure the "influence" of a scientific article, is based on the following three phases, by applying some experimental results of [8]:

1. **Subgraph building**: Starting form certain target nodes (articles), for which we are interested in measuring their scientific impact, and expanding backward by following reversely the nodes having out-going links to the target nodes. The procedure stops after a fixed number of levels. This can be done by an iterative deepening depth-first search. In this work, the graphs contain all nodes, from which the target nodes can be reached in at most three steps and we consider the *induced subgraph* of that nodes.

2. **Estimating the PR of the boundary**: We use a heuristic to estimate the individual $PR$: in each iteration turn, we add an extra term to the $PR$ value of each boundary node that equals to the fraction of its in-coming edges to all edges in the subgraph.

3. **Calculating the $PR$ and $RP$**: On one hand, we run the PageRank algorithm on the subgraph, in each step we use the estimated $PR$ value of the boundary nodes adding the $\lambda/N$ damping factor to each node. On the other hand, we also calculate the reaching probability, $RP$, of the target node(s) in the subgraph.

The idea behind the necessity of the second phase is that, although the PageRank values cannot be calculated exactly without having run the algorithm on the full graph, still the estimation heuristic we defined gives an acceptable approximation for the constructed subgraph as it has been already proven in [2], and tested by simulations [8]. We also note that the convergence of the PageRank is guaranteed by this method opposite to that one defined by Csendes and Antal for the same

**Algorithm 1**: Local PageRank method for a scientific article

```
Input  : Scientific article ID $A$.
Output: The $PR$-score of the article from its local co-citation network,
Build the article's local co-citation network with radius $r$
Fix the PageRank values of each boundary node $v$ as $PR(v) = |N^+(v)|/|E(G)|$
Calculate $PR$-scores of each node in the subgraph by using the PageRank algorithm
Return $PR(A)$.
```

purpose. We set the radius size $r = 3$ from the target nodes because of two reasons: the first is that the number of nodes in the fourth layer is $\mathcal{O}(N)$ and the in-degrees are bounded with a constant, thus, with respect to PageRank algorithm, it is enough to consider the number of in-coming links to the boundary nodes from this layer, and not to consider the linking structure between them to get a good approximation of the $PR$-scores. The second reason is that we assume, that the articles at a distance more than three (with respect to the co-citation graph) do not have much impact on the target articles in scientific sense (which may be acceptable in scientometrics).

## 4. Results and discussion

As it is known, Harold Kuhn developed an algorithm for solving the assignment problem [18] and he named it as the Hungarian method acknowledging the contribution of Jenő Egerváry and Dénes Kőnig [10, 17]. The paper of Egerváry received just a few citations (probably because it was written in Hungarian) while some of the citing papers received much more: for Egerváry's paper 38 citations can be found in the ISI Web of Knowledge database, while the artice of Kőnig and Kuhn received there 215 and 726, respectively. In contrast to classic scientometrics that only takes into account the direct number of citations, we shall see that the network based methods show a more realistic picture of the importance of Egerváry's paper.

We constructed a network which contains the following articles as nodes: the famous paper of Jenő Egerváry: *On combinatorial properties of matrices* (published in Hungarian, 1931), the three articles which referred in Egerváry's paper, the articles that cite Egerváry's one, all articles that cite at least one of the previous ones and all articles that cite articles on the "second level". We consider the network that is induced by these nodes as described in the first phase; it contains $N = 1155$ nodes and 1923 edges. Figure 1 shows the network, where the paper of Egerváry highlighted with big black square.

We applied the modified PageRank algorithm (with $\lambda = 0.1, 0.15, 0.2, 0.25$) described in Section 3 for this network and also calculate the reaching probabilities of the nodes. We observed that the PageRank method is robust against the choice of $\lambda$. The results (with $\lambda = 0.2$) are summarized in Table 1 for four notable publications in the co-citation network.

| Publication | $PR$-Score | $PR$-rank | $RP$-score | $RP$-rank | #Cites | Cite rank |
|---|---|---|---|---|---|---|
| Egervári [10] | 0.891 | 4 | 0.009 | 2 | 39 | 65 |
| Kuhn [18] | 1.189 | 1 | 0.042 | 1 | 726 | 1 |
| Ford, Fulkerson [13] | 0.525 | 8 | 0.004 | 9 | 39 | 65 |
| Bellman [3] | 0.399 | 11 | 0.003 | 10 | 18 | 158 |

Table 1: PR-score (with $\lambda = 0.2$), reaching probabilies and number of citations of the famous publications in the Egerváry co-citation graph. PR-score is multiplied by $10^2$

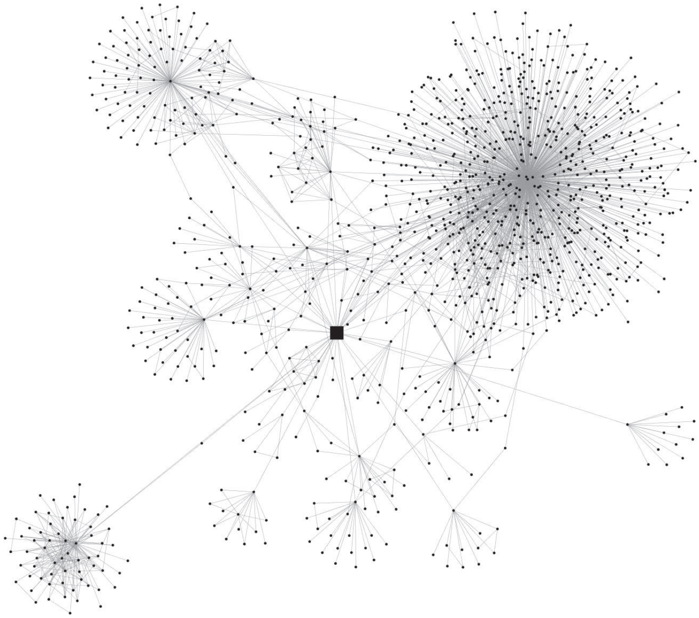

Figure 1: Local co-citation network containing the famous paper of Egerváry (highlighted with big square)

First, we observed that the choice of the damping factor $\lambda$ does not influence the final ranking of the first ten publications, only small changes can be noticed in the rest of the ranking. The ranks and the relative values of the papers to each other show a more realistic picture of the importance of them. It is not surprising, that Kuhn's paper $PR$ value is the highest by far, the 726 citations for this paper is outstanding in the field. The second and third articles in the $PR$ rank became D. Kőnig: *Graphs and their applications for the theory of determinants and sets* (in Hungarian, 215 citations) and G. Frobenius : *Über zerlegbare Determinanten* (11 citation), respectively. Both articles were cited in Egerváry's paper which became the fourth highest ranked paper although it only received 39 citations and that it is only in the 65th place in the citation ranking. The very high position of Forbenius's paper in the ranking is definitely due the reputation it obtains from Egerváry's article. It is worth highlighting that Ford and Fulkerson's article, which received the same number of citations as that of Egerváry, was ranked lower but it is still in the top ten. This two facts also indicate the advantages of the PageRank based evaluation, since this paper was also quite important in the development of operation research. We also point out, that the similarly important paper of Bellman was ranked 11th (although it received just 18 citations) which shows a much clearer picture of its impact (in contrast to its citation rank). It is also interesting to observe, that the $RP$-rank of Egerváry's article is two, which means that a random searcher who checks the articles of the field finds that paper with the second highest probability.

We hope that network-based ranking methods gain more space in scientometric since they show a more objective picture of the impact of scientific publications. It follows from the implementation of the PageRank algorithm that citations received from more important papers contribute more to the ranking of the cited paper than those coming from less important ones. Furthermore, simplicity and fast computability of this method are also advantageous. On the other hand, co-citation networks give a more detailed contextual information (compared to the number of citations) for evaluating the impact of an article.

**Acknowledgment** The authors are grateful to Elvira Antal for constructing the network and for her respective contribution.

# References

[1] Alonso, S., Cabrerizo, F., Herrera-Viedma, E., and Herrera, F. H-index: A review focused in its variants, computation and standardization for different scientific fields. *Journal of Informetrics 3*, 4 (2009), 273–289.

[2] Bar-Yossef, Z., and Mashiach, L.-T. Local approximation of pagerank and reverse pagerank. In *Proceedings of the 17th ACM conference on Information and knowledge management* (2008), ACM, pp. 279–288.

[3] Bellman, R. Mathematical aspects of scheduling theory. *Journal of the Society for Industrial & Applied Mathematics 4*, 3 (1956), 168–205.

[4] Bergstrom, C., West, J., and Wiseman, M. The eigenfactor metrics. *The Journal of Neuroscience 28*, 45 (2008), 11433–11434.

[5] Brin, S., and Page, L. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems 30*, 1 (1998), 107–117.

[6] Chen, C. Visualising semantic spaces and author co-citation networks in digital libraries. *Information Processing & Management 35*, 3 (1999), 401–420.

[7] Chen, P., Xie, H., Maslov, S., and Redner, S. Finding scientific gems with google's pagerank algorithm. *Journal of Informetrics 1*, 1 (2007), 8–15.

[8] Chen, Y.-Y., Gan, Q., and Suel, T. Local methods for estimating pagerank values. In *Proceedings of the 13th ACM International conference on Information and knowledge management* (2004), pp. 381–389.

[9] Csendes, T., and Antal, E. Pagerank based network algorithms for weighted graphs with applications to wine tasting and scientometrics. In *Proceedings of the 8th International Conference on Applied Informatics* (2010), pp. 209–216.

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

[11] Egghe, L. An improvement of the h-index: The g-index. *ISSI Newsletter 2*, 1 (2006), 8–9.

[12] Fiala, D., Rousselot, F., and Ježek, K. Pagerank for bibliographic networks. *Scientometrics 76*, 1 (2008), 135–158.

[13] Ford, L. R., and Fulkerson, D. Solving the transportation problem. *Management Science 3*, 1 (1956), 24–32.

[14] Hirsch, J. An index to quantify an individual's scientific research output. In *Proceedings of the National Academy of Sciences of the USA* (2005), vol. 102, pp. 16569–16572.

[15] Jeong, H., Néda, Z., and Barabási, A. Measuring preferential attachment in evolving networks. *Europhysics Letters 61*, 4 (2007), 567–572.

[16] Kleinberg, J. Authoritative sources in a hyperlinked environment. *Journal of the ACM 46*, 5 (1999), 604–632.

[17] König, D. Über graphen und ihre anwendung auf determinantentheorie und mengenlehre. *Mathematische Annalen 77*, 4 (1916), 453–465.

[18] Kuhn, H. W. The hungarian method for the assignment problem. *Naval research logistics quarterly 2*, 1-2 (1955), 83–97.

[19] Kumar, M. Evaluating scientists: Citations, Impact factor, h-index, Online page Hits and What Else? *IETE Technical Review 26*, 3 (2009), 165–168.

[20] Lehmann, S., Lautrup, B., and Jackson, A. Citation networks in high energy physics. *Physical Review E 68*, 2 (2003), 026113.

[21] Liu, X., Bollen, J., Nelson, M., and Van de Sompel, H. Co-authorship networks in the digital library research community. *Information processing & management 41*, 6 (2005), 1462–1480.

[22] Norris, J. R. *Markov chains*. Cambridge University Press, 1998.

[23] Radicchi, F., Fortunato, S., Markines, B., and Vespignani, A. Diffusion of scientific credits and the ranking of scientists. *Physical Review E 80*, 5 (2009), 056103.

[24] Su, C., Pan, Y., Zhen, Y., Ma, Z., Yuan, J., Guo, H., Yu, Z., Ma, C., and Wu, Y. Prestigerank: A new evaluation method for papers and journals. *Journal of Informetrics 5*, 1 (2011), 1–13.

[25] Walker, D., Xie, H., Yan, K., and Maslov, S. Ranking scientific publications using a model of network traffic. *Journal of Statistical Mechanics: Theory and Experiment 2007*, 06 (2007), P06010.

[26] Wendl, M. C. H-index: however ranked, citations need context. *Nature 449* (2007), 403.

[27] Woeginger, G. An axiomatic characterization of the Hirsch-index. *Mathematical Social Sciences 56*, 2 (2008), 224–232.

[28] Yan, E., and Ding, Y. Discovering author impact: A pagerank perspective. *Information Processing & Management 47*, 1 (2011), 125–134.