Exploration and Evaluation of Citation Networks

Karel Jezek[1]; Dalibor Fiala[2]; Josef Steinberger[1]

[1]Department of Computer Science & Engineering, University of West Bohemia
Univerzitni 8, 306 14 Pilsen, Czech Republic
e-mail: jezek_ka@kiv.zcu.cz; jstein@kiv.zcu.cz
[2]Gefasoft AG
Dessauerstrasse 15, 80992 Munich, Germany
e-mail: dalibor.fiala@gefasoft.de

Abstract

This paper deals with the definitions, explanations and testing of the PageRank formula modified and adapted for bibliographic networks. Our modifications of PageRank take into account not only the citations but also the co-authorship relationships. We verified the capabilities of the developed algorithms by applying them to the data from the DBLP digital library and subsequently by comparing the resulting ranks of the sixteen winners of the ACM SIGMOD E.F.Codd Innovations Award from the years 1992 till 2007. Such ranking, which is based on both the citation and co-authorship information, gives better and more fair-minded results than the standard PageRank gives. The proposed method is able to reduce the influence of citation loops and gives the opportunity for farther improvements e.g. introducing temporal views into the citations evaluating algorithms.

Keywords: WWW structure mining; PageRank; citation analysis; citation networks; ranking algorithms; social networks;

1. Introduction

Rating of research institutions and researchers themselves is a challenging and important area of investigation. Its conclusions have a direct influence on acquiring financial support for research groups. The aim of our work is to investigate citation networks (networks of relationships between citing and cited publications) and other similar networks, e.g. hyperlink structures of the Web. We want to derive a rating of individual participants modeled as nodes of the network graph.

Every system modeled as a graph is a network. These two notions are actually synonyms. Perhaps the word graph has a more abstract meaning and therefore mathematicians prefer speaking of graphs rather than networks which are the notion in the terminology of social sciences.

Real world networks are grouped into social, information, technological and biological networks [1]. Citation networks as well as Word Wide Web hyperlink structures are mostly included in information networks, but some authors [2] use the term "social" in this context. As stated above, network systems can be modeled as a graph. Mathematical notions and formulas from graph theory are available to explore their features and results from one type of networks are profitably utilized in others.

In Section 2 we are concerned with ranking of Web pages. Methods originated for determination of page importance were soon recognized as applicable to citation analysis. Connections between the ranking method and co-authorship networks are discussed in Section 3. Section 4 is the core of the article and introduces our evaluation method of citation networks. The next part presents results achieved on data from the DBLP digital library. Finally, possible further improvements are proposed together with other

application areas where the introduced method can be used.

2. Ranking of Network Structures

WWW is a gigantic extensive explored network structure. Filtering Web documents by relevance to the topic the user is interested in does not sufficiently reduce the number of searched documents. Some further criteria must decide which documents are worth the user's attention and which are not. In [3] Page and Brin proposed an iterative calculated page ranking (or topic distillation) algorithm based on hyperlinks. This algorithm, suitably named PageRank, has at the same time been used in the famous search engine Google, and without doubt it is one of the basic reasons behind Google's successes. The PageRank technique is able to order Web documents by their significance. Its principle lies in collecting and distributing "weights of importance" among pages according to their hyperlink connections. Figure 1 demonstrates PageRank calculations for a piece of a hypothetical network. It assigns high ranks to pages that are linked to by documents that themselves have a high rank. The whole process runs iteratively and represents probably the world's largest matrix computation.

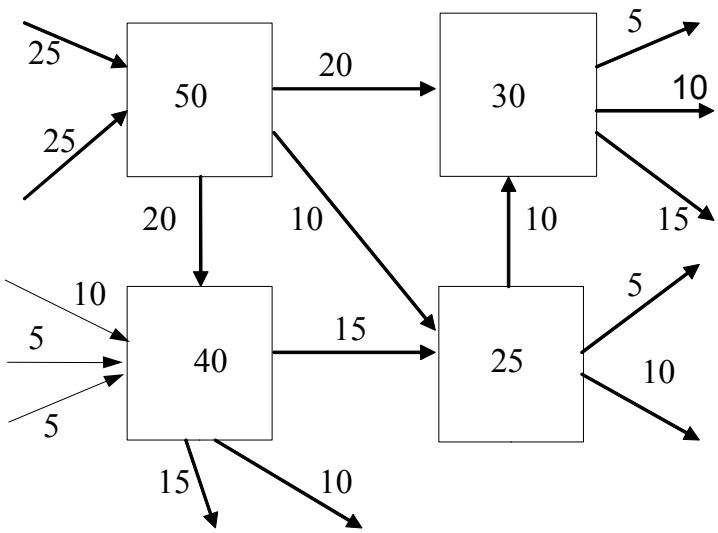

Figure 1: Rank distribution and collection within a PageRank calculation

Approximately at the same time as PageRank appeared, J. Kleinberg [4] proposed a similar algorithm for determining significant web pages called HITS. Other new ranking methods and modifications soon appeared - SALSA, SCEAS Rank, ObjectRank, BackRank, AuthorRank, etc. To prove the applicability of a method for rating research institutions, we collected the Web pages of main Czech computer science departments and applied the rating formula to their hyperlink structure [5].

2.1 PageRank

Let us briefly introduce the PageRank principles as presented in [3] and [6]. Let $G = (V,E)$ be a directed graph, where V is a set of vertices (corresponding to Web pages) and E a set of edges (representing hyperlinks between Web pages). The PageRank score $PR(u)$ for Web page u is defined as:

$$PR(u) = \frac{1-d}{|V|} + d \sum_{(v,u)\in E} \frac{PR(v)}{D_{out}(v)} \qquad (1)$$

where $|V|$ is the number of nodes, d is the dumping factor (an empirically determined constant set between 0.8 and 0.9) and $D_{out}(v)$ is the out-degree of node v (number of outgoing edges from node v). You can see that the PageRanks of nodes depend on the PageRanks of other nodes. As the hyperlink structure is usually cyclic, so the PageRank evaluation is a recursive process allowing the current node to influence all nodes to which exists the path from the current node.

The randomizing factor *(1-d)* represents the possibility to jump to a random node in the graph regardless of the out-edges from the current node. On the contrary, *d* stands for the probability of following out-link from the present node. Introducing the random term prevents loops of nodes (rank sinks) from accumulating too much rank and not propagating it further. An example of a rank sink is illustrated in Figure 2. There are also problems with nodes without out-links (referred to as dangling pages in PageRank evaluation) that would not distribute their rank either. In fact, zero-out-degree Web pages and rank sinks are the main problems in PageRank processing. On the other hand, nodes without in-links are not harmful and their rank is always smaller than that of any nodes with some in-links, as expected.

The PageRank method is rather reliable. The necessary number of iterations depends on the extensiveness of the Web graph, but converges promptly. For a graph with over 320 million nodes (pages), only about 50 iterations were required as claims [3]. The frequency of normalization and the order of nodes affect the final ranking, but the effect on the resulting rank is not substantial.

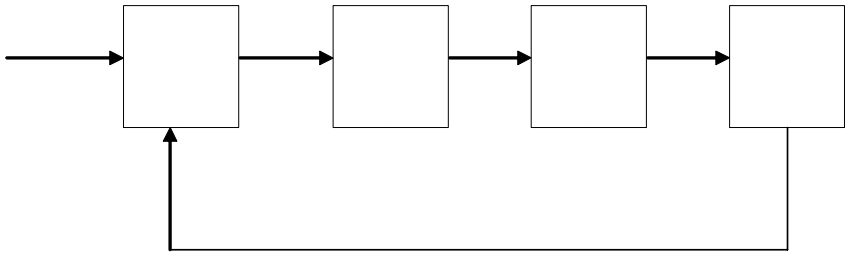

Figure 2: An example of a graph with a rank sink

We evaluate an iterative calculation of PageRank as follows:

1. We remove duplicate links and self-links from the graph.
2. We set the initial PageRanks of all nodes in the graph uniformly so that the total rank in the system is one. This is the zeroth iteration.
3. We remove nodes having no out-links iteratively because removing one zero-out-degree node may cause another one to appear.
4. We compute the PageRank scores for all nodes in the residual graph according to Figure 1, using the scores from the previous iteration. We perform normalization so that the total rank in the system (including the vertices removed in step 3) is again one.
5. We repeat step 4 until convergence. Numerical convergence of the scores is usually not necessary. An ordering of nodes (by PageRank) that does not change (or changes relatively little) is satisfactory as claims [7].
6. We gradually add back the nodes removed in step 3, compute their rank score and re-normalize the whole system.

Normalization of the rank obtained from in-linking nodes by their out-degree is an important feature of PageRank. In this way, such nodes are penalized which are connected to many other nodes. It corresponds to a similar situation in citation evaluation, when citations of frequently citing authors are less valuable then those citing rarely. This analogy was a motivating idea for applying PageRank principles to bibliographic

citations.

2.2 SCEAS Rank

In [8] an iterative PageRank like the SCEAS method (Scientific Collection Evaluator with Advanced Scoring) is used to rank scientific publications. It evaluates the impact of publications on the basis of their citations. In the graph where nodes are publications and edges mean citations between them, the original PageRank metrics is not appropriate. Such graph often contains cycles which are in fact a kind of self-citation. Therefore, we would rather the nodes from the cycle not have much influence on rank distribution. Similarly, the direct citations should have their impact higher than indirect citations and their impact should become smaller when the distance between cited and citing gets larger.

$$R(u) = (1-d) + d \sum_{(v,u)\in E} \frac{R(v)+b}{D_{out}(v)} a^{-1} \qquad \text{where } (a \geq 1,\ b>0) \tag{2}$$

The SCEAS formula (2) computes the rank score *R(u)* with direct citation enforcing factor *b* and speed *a* in which an indirect citation enforcement converges to zero. For *b=0* and *a=1* formula (2) is equivalent to PageRank formula (1). The authors experimentally proved that SCEAS converges faster than PageRank. They carried out experiments with data from the DBLP digital library and compared the SCEAS rankings with several other ranking schemes including PageRank, HITS and a "baseline" ranking constituted of authors winning an ACM award. They showed that their method is superior to the others. We adopted their comparison methodology to test our novel algorithm.

2.3 Other ranking methods

As mentioned above, PageRank is not the only method of ranking. The most elementary way is to count in-links for each node. The most authoritative node is then the one with the highest number of in-linking edges. The rank $R_{in}(u)$ of node *u* can be computed as:

$$R_{in}(u) = \sum_{(v,u)\in E} w(v,u) \tag{3}$$

In the case in which the graph *G* is unweighted, e.g. all weights *w(v,u)* are equal to one, the sum of in-linking edges gives an in-degree of the node. If applied to citations, all have the same weights and the citation of B in A does not influence the citation of C in B. Publication C is in (3) ranked as if it was not indirectly (through B) cited in A. Note that PageRank preserves such transitive feature respecting contributions of reputation from outlying nodes.

Another ranking technique worth mentioning is HITS [4], [9]. HITS (Hyperlink-Induced Topic Search) defines two values (authority *A(u)* and hubness *H(u)*) for each node *u* as follows:

$$\begin{aligned} A(u) &= \sum_{(v,u)\in E} H(v) \\ H(u) &= \sum_{(v,u)\in E} A(v) \end{aligned} \tag{4}$$

Importance of the node has two measures. The nodes pointed to by many nodes with high hub scores have high authorities and the nodes pointed to by many good authorities have high hubness. Mutual reinforcement between hubs and authorities is evident. HITS is applicable to citation networks as well and

gives reasonable results. The necessity to work with two scores was the main reason why we preferred the PageRank algorithm for our further research.

A simple metric of researcher scoring called the *h-score* was proposed in [10]. A researcher has a score *h* if *h* of his papers have at least *h* citations each. The *h-score* enables you to evaluate the successfulness of researchers at different levels of seniority. When *n* is the number of years in service of a researcher (since the year of his first publication), then his successfulness *m* is calculable as:

$$m \approx h / n \tag{5}$$

E.g. a scientist in physics is successful if his/her *m* is close to 1. The *h-index* has obvious advantages. It is only a single number; it does not prefer quantity to quality. On the other hand it is not comparable across different scientific fields and does not reflect co-authorships.

3. Co-authorship networks and Ranking Methods

Co-authorship networks are a special case of social networks, in which the nodes represent authors and edges mean collaboration between authors. Unlike the citation networks mentioned above, in which each edge means acknowledgement of primacy, declaration of debt or recognition, in a co-authorship graph an edge connecting two authors expresses the fact that those authors are or were colleagues. They have published one or more articles as a result of common research lasting for a year or years. This is in contrast to such citations where the citing author does not know the cited author personally and these persons have never collaborated. Co-authorship networks can also express the intensity of cooperation. We can consider a number of co-authors in the paper or a number of common papers to assess the weight of cooperation.

3.1 AuthorRank method

A co-authorship network model is investigated in [2]. It introduces AuthorRank as an indicator of the importance of an individual author in the network. As the number of collaborated authors is rather limited, the co-authorship graph of all documents consists of strongly connected components whose number may be huge but can be evaluated independently. The AuthorRank result gives the impact scores of authors using similar principles to PageRank. Let us briefly mention the main idea of AuthorRank.

Any co-authorship network can be described simply as an undirected unweighted graph, where nodes represent authors and edges symbolize the existence of collaboration. If we allow a variety of authorities in the graph, we have to replace any undirected edge between nodes e.g. a_1 and a_2 with two directed edges (one directed from a_1 to a_2 , the second directed from a_2 to a_1). Further, we have to weight the collaboration not uniformly, e.g. assign weights w_{ij} to edges. Therefore, we need some additional knowledge which is not included in the undirected co-authorship graph. To show the case in a non-trivial but simple enough example, let us suppose as in [2] three cooperating authors. Figure 3 shows their co-authorship graphs.

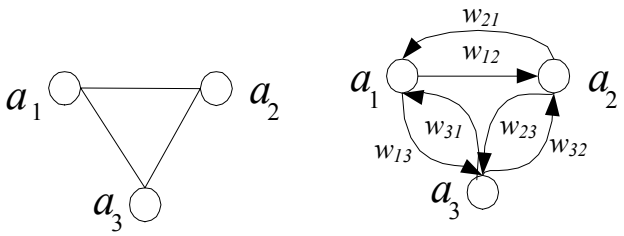

Figure 3: Co-authorship graph

The remaining but substantial problem is determination of weights *w*. Co-authors of a paper published by two authors are obviously more tightly connected than co-authors of a paper written by ten authors. Frequently collaborating authors should be more connected than the authors jointly publishing only occasionally.

This problem is solved in [2] with the help of two factors used in the collaboration graph – co-authorship frequency and exclusivity. They should give higher weight to edges that connect authors often publishing together with a minimum number of other authors involved.

Let m be the number of publications, N the number of authors and $f(p_k)$ the number of authors of publication p_k. Then co-authorship exclusivity $g_{i,j,k}$, frequency c_{ij} and on their basis evaluable weight w_{ij} (between authors a_i and a_j) can be computed following way:

$$\begin{aligned} g_{i,j,k} &= 1/(f(p_k)-1) \\ c_{ij} &= \sum_{k=1}^{m} g_{i,j,k} \\ w_{ij} &= \frac{c_{ij}}{\sum_{k=1}^{N} c_{ik}} \end{aligned} \qquad (6)$$

The weights are normalized (divided by the sum of weights of outgoing edges from the node), which is necessary for convergence of an algorithm computing nodes' prestige. The resulting AuthorRank of an author i is evaluated as follows:

$$AR(i) = (1-d) + d \sum_{j=1}^{N} AR(j) \times w_{ij} \qquad (7)$$

where $AR(j)$ corresponds to the AuthorRank of node j from which goes the edge to node i with weight w_{ij}.

Let us remember that the above described method works with collaborations not with citations. We believe that to measure the importance or prestige of nodes only on the basis of collaboration is questionable at least. Why should researchers who have many co-authors be more authoritative than those having just a few co-workers? Consider e.g. authors frequently publishing their works without co-authors. They are strongly handicapped in the AuthorRank methodology and completely ignored in the extreme case – publishing without co-authors at all. Single-author papers are quite common. In the DBLP collection we used in our experiments they made up 1/3 of them. The authoritativeness in the collaboration networks does not reflect the authoritativeness based on citations. But just citations are an accepted means of evaluating a researcher's importance.

4. Citation analysis and co-authorship

The main objective of this article is adapting the PageRank method to the citation analysis task. There are other PageRank modifications, e.g. the one submitted by [8] is meant particularly for bibliographic citations. The original contributions of our work are extensions and improvements of a traditional citation analysis method. Our innovations are based on considering mutual cooperation between the cited and citing author and its various assessments. If we allow the existence of co-authorship influence on citations, we might want to refine the citation analysis results. To consider the higher impact of a citation between not cooperating authors, we need to involve co-authorship networks in the evaluation process.

Our rating model is based on three graphs which are all derivable from digital library documents. This model includes:

i) bipartite graph of co-authorship,
ii) publication-citation graph,
iii) author-citation graph.

A simple example of graphs is shown in Figure 4.

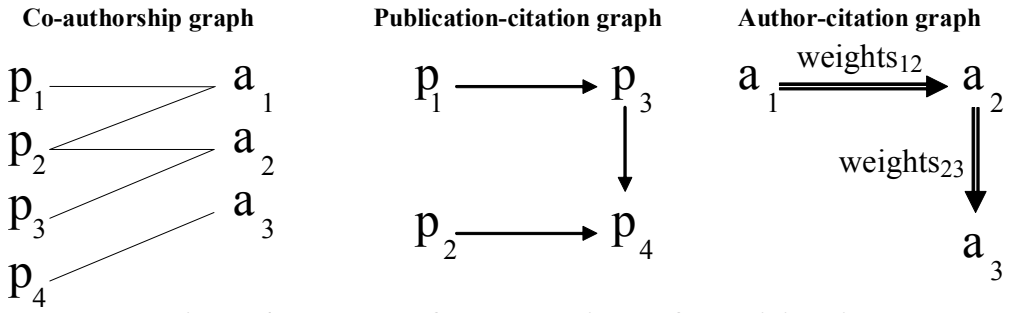

Figure 4: Example of graphs derivable from digital library

Ad i:

The nodes of this unweighted graph consist of two disjunctive sets. One contains authors and the second publications. The edges are undirected matching authors and their publications.

Ad ii:

This graph is unweighted and its nodes represent publications. The edges are directed and express bindings between citing and cited publications. No common authors in a citing and cited publication are allowed.

Ad iii:

It is an edge-weighted directed graph. Its nodes represent the set of authors. Edges represent the citation between the authors. This graph is derivable from those two mentioned above. A triple (w_{uv} , c_{uv} , b_{uv}) of weight is associated with each edge, where: w_{uv} represents the number of citations between citing author u and cited author v, c_{uv} is the number of common publication by authors connected with this edge, b_{uv} expresses various semantics of collaboration we want to stress. E.g. the overall number of publication of both authors, the overall number of co-authors, the overall number of distinct co-authors and some other alternatives giving a true picture of the cooperation effect on citations. Actually, the author-citation graph should have the form of a multi-graph and the introduced triples substitute the multiplicity of its edges.

For those who prefer mathematical symbolism let us define the above introduced graphs formally. It allows us to exactly express the weights assigned to the edges of the author-citation graph:

i. The co-authorship graph $G^P = (P \cup A, E^P)$ is an undirected, unweighted, bipartite graph, where $P \cup A$ is a set of vertices (P a set of publications, A a set of authors) and E^P is a set of edges. Each edge $(p, a) \in E^P$, $p \in P$, $a \in A$ means that author a has co-authored publication p.
ii. The publication-citation graph $G^C = (P, E^C)$ is a directed unweighted graph, where P is a set of vertices representing the publications, and E^C is a set of edges. The edge $(p_i, p_j) \in E^C$ denotes a citation of publication p_j in publication p_i.
iii. The author-citation graph $G = (A, E)$ is a directed, edge-weighted graph, where A is a set

of vertices representing authors and E is a set of edges denoting citations between authors. For every $p \in P$ let $A_p = \{a \in A: \exists(p, a) \in E^P\}$ be the set of authors of publication p. For each (a_i, a_j), $a_i \in A$, $a_j \in A$, $a_i \neq a_j$, where exists $(p_k, p_l) \in E^C$ such that $(p_k, a_i) \in E^P$ and $(p_l, a_j) \in E^P$ and $A_{pk} \cap A_{pl} = \emptyset$ (i.e. no common authors in the citing and cited publications are allowed) there is an edge $(a_i, a_j) \in E$. Thus, $(a_i, a_j) \in E$ if and only if $\exists(p_k, p_l) \in E^C \wedge \exists(p_k, a_i) \in E^P \wedge \exists(p_l, a_j) \in E^P \wedge A_{pk} \cap A_{pl} = \emptyset \wedge a_i \neq a_j$.

The weight $w_{u,v}$ representing the number of citations from u to v can now be defined as:

$$w_{u,v} = |C|, \text{ where } C = \{p_k \in P: \exists(p_k, u) \in E^P \wedge \exists(p_l v) \in E^P \wedge \exists(p_k, p_l) \in E^C \wedge p_k \neq p_l\}.$$

The weight $c_{u,v}$ representing the number of common publications by u and v is defined as:

$$c_{u,v} = |CP|, \text{ where } CP = \{p \in P: \exists(p,u) \in E^P \wedge \exists(p,v) \in E^P\}.$$

The third weight $b_{u,v}$ symbolizes the values obtained from the various formulas we have used in our experiments. They should more softly express the examined views of the author's cooperation. The considered alternatives were:

a. $b_{u,v} = |P_u| + |P_v|$ where $P_i = \{p \in P: \exists (p, i) \in E^P\}$, e.i. the total number of publications by u plus the total number of publications by v,

b. $b_{u,v} = |ADC_u| + |ADC_v|$ where $ADC_i = \{a \in A: \exists p \in P$ such that $(p, a) \in E^P \wedge (p, i) \in E^P\}$, i.e. the number of all distinct co-authors of u plus the number of all distinct co-authors of v,

c. $b_{u,v} = |ADC_u| + |ADC_v|$ where ADC_i is defined as above but it is a multiset, i.e. the number of all co-authors of u plus the number of all co-authors of v,

d. $b_{u,v} = |DCA|$ where $DCA = \{a \in A: \exists p \in P$ such that $(p, a) \in E^P \wedge (p, u) \in E^P \wedge (p, v) \in E^P\}$, e.i. the number of distinct co-authors in common publications by u and v,

e. $b_{u,v} = |DCA|$ where DCA is defined as above but it is a multiset, i.e. the number of co-authors in common publications by u and v,

f. $b_{u,v} = |P_u| + |P_v| - |SP_u| - |SP_v|$ where $P_i = \{p \in P: \exists (p, i) \in E^P\}$ and $SP_i = \{p \in P:(p,i) \in E^P \wedge d_{G^P}(p) = 1\}$, i.e. the number of publications by u where u is not the only author plus the number of publications by v where v is not the only author,

g. $b_{u,v} = 0$, i.e. no refinements by $b_{u,v}$ are introduced.

The weights are used as parameters in a modified PageRank formula (see below), where the main innovative part is a function of $w_{u,v}$, $c_{u,v}$, $b_{u,v}$, named *contribution*(u, v) and used as a multiplicative factor of the contributing ranks. The rank of each author u evaluates from ranks of him citing authors (there exists the edge (u,v) from the citing author u to the cited author v). The rank formula is not as complicated

$$R(v) = \frac{1-d}{|A|} + d \sum_{(u,v) \in E} R(u) \frac{contribution(u,v)}{\sum_{(u,k) \in E} contribution(u,k)} \qquad (8)$$

as it looks at first sight; its similarity with the original PageRank is evident.

Except for *contribution* the meaning of other symbols was explained above; the rank of cited author v is counted from the rank of him citing author u, d is as usual the dumping factor, an empirically determined constant set to 0.85. The contribution from u to v must be normalized (divided by the sum of contributions from u). The sum of all contributions must be 1 to guarantee convergence. The *contribution(x,y)* is evaluated by formulas (9).

$$contribution(x, y) = \frac{w_{x,y}}{f(c_{x,y}, b_{x,y}) \sum_{(x,j)\in E} w_{x,j}}$$

where (9)

$$f(c_{x,y}, b_{x,y}) = \frac{c_{x,y} + 1}{b_{x,y} + 1}$$

The goal of the presented modification is to penalize the cited authors if they frequently collaborate with the citing authors. The *contribution*(*x, y*) defined in (9) on the one hand increases prestige of the node *v* in formula (8) proportionally to the number of its citations but on the other hand it reduces its prestige when the citing author has published (some other publication) together with the cited (see $c_{x,y}$ in $f(c_{x,y}, b_{x,y})$). The reduction was again chosen as proportional to the number of (common) publications. The tightness of binding between the citing and cited author when they together published some other papers (note that no common authors in citing and cited publications are allowed) should depend on the number of their co-authors. Therefore, we introduced the term $b_{x,y}$ in the formula. Its variations were mentioned above inclusive of the zero value discarding its effect. The constant 1 is used to prevent zero dividing and the sum of $w_{x,j}$ is for normalization. Roughly speaking, *contribution*(*x, y*) represents the normalized weight of citations from *x* to *y* with respect to the author's cooperation.

In case authors *x* and *y* have no common publications, the coefficient $c_{x,y}$ is zero, $b_{x,y}$ is then implicitly zero in the alternatives d, e and according to the definition in the alternative g. The other alternatives assigning the $b_{x,y}$ value on the basis of the total number of author's publications or co-authors in the environment where any common publications *x* and *y* does not exist should due to the definitions be non-zero. But this non-zero value is not justifiable. There is no reason to contribute to the author's rank from one citation more or less depending on the total number of his publications or co-authors. Therefore, whenever $c_{x,y}$ is zero we assign to $b_{x,y}$ zero too. When the coefficients $c_{x,y}$ and $b_{x,y}$ are all zero, formula (8) corresponds to the weighted PageRank used e.g. in [11].

Certainly it is possible to deduce other formulas to express the influence of the author's cooperation on the citation. The method just described works well, as we will show in the next section. Other alternatives and experiments will be investigated in the future.

5. Evaluation

We tested our formula for various alternatives of the function of $w_{u,v}$, $c_{u,v}$, $b_{u,v}$ on a bibliographic dataset derived from the DBLP library available in XML format. The *http://dblp.uni-trier.de/xml/dblp20040213.xml.gz* version of the collection was used. Only journal and proceedings papers similar to [8] were extracted. Nearly half a million journal and conference papers were explored. Over eight thousand of them have references, but some of them are outside the DBLP library.

The investigated publication-citation graph has approximately five hundred thousand nodes and around one hundred thousand edges. The derived co-authorship graph was much wider, with around eight hundred thousand nodes (authors + publications) and one million edges, each of them representing an author-publication couple. The most frequent number of co-authors is two, an average is 2.27. The relevant author-citation graph contains over three hundred thousand nodes and nearly the same number of edges. Fifteen thousand authors were not isolated.

There is a problem of how the ranking method should be assessed. The author's prestige surely depends on citations, but there are many choices, as stated above. Our results should reflect a common human

meaning. They should approximate the meaning of a broad group of professionals in the rating domain. Therefore, we decided to use approved ACM honors. The resulting ranks were compared by sixteen winners of the SIGMOD E. F. Codd Innovation Awards from the years 1992 till 2007. We supposed the rank of winners should be relatively high and the positions of winners provide an evaluation of the abilities of the used formulas.

6. Results

The rankings received by our modified formula were clearly better (relative to the Codd Award winners) than those received by the standard PageRank. The sum of ranks, the worst rank and the median rank of winners were used as indicators of rating quality. The "outlierless" median omits the worst column value. Table 1 presents the results.

There is a drawback when a time sequence of award-winners is used for quality ranking evaluation. The "oldest" award-winners, as you can see in Table 1, occupy the best positions in all columns. It is explainable as "the permanency effect"; they take advantage of their popularity, i.e. becoming more popular and prestigious, they are more often cited.

The column labeled "PageRank" shows the results of the standard PageRank formula and serves as a baseline. The next column gives results when the weighted PageRank is used. Remarkable improvements are obvious. The next seven columns present the results of modifications a – g of formula (9). The best behavior is seen in the b and c columns. It confirms the last row too, showing the median rank when the worst place is disregarded. This is a common practice when an outlier can distort the data. The last two columns are just for reference. The relatively simple "In degree" behaves well and "HITS authorities" in the last column surprisingly significantly overcome the basic PageRanks.

7. Conclusion

Graph theory is a traditional discipline originating from the eighteen century. Its utilization in information network analysis is only a few years old and is being intensively investigated with the expansion of the Web. Novel methods developed initially for Web mining were recognized as useful and applicable in citation analysis as well. This contribution presented an overview of the most important and recent methods from the field of Web pages, articles and author citation analysis. We concentrated on the issue of analyzing the network structure in order to find authoritative nodes. The main contributions of our work are modifications of the PageRank equation, this time suited for graphs of citations between publications and collaborations between authors. This enables one to rank authors "more fairly" by significance, taking into account not only citations but also collaborations between them.

To test this new approach on actual data, we applied our ranking algorithms to a data set from the DBLP digital library and used the methodology of Sidiropoulos and Manolopoulos [8] for ranking comparisons. We compared author rankings to a list of ACM SIGMOD E. F. Codd Innovations Award winners and found that the new rankings much better reflected the prize award scheme than the baseline, "standard" PageRank ranking. It was not possible to directly compare our results with those of Sidiropoulos et al. because they utilized a slightly different data set and their method is primarily destined for publications, not for authors.

Our experiments proved that adding the aspect of the author's cooperation to the ranking algorithm improves the rating performance. Nowadays, large electronic libraries give the best chance of ranking scholars, research groups or even whole institutions - from departments to universities.

There are many exciting research directions in the areas of bibliometrics, webometrics and scientometrics.

In future research, we plan to continue primarily in the following directions:

- It seems to be useful to more carefully analyze the sensitivity and stability of computations on parameters *b, c, w* in formulas (8), (9). Our next aim has to be their more expedient integration into the ranking formula. This presently used is based only on simple reasoning. Although the standard PageRank has been shown to be relatively stable, the larger number of parameters involved in the calculation may negatively affect this property.
- We expect further improvements and more fair-minded results when time relations between citing and cited items will be included in the ranking evaluation. Time stamps are or at least should be an ordinary part of bibliographical records and they may certainly be beneficially utilized. The concept of a "fairer" ranking of researchers based not only on citations but also on collaborations invites inclusion of the time factor. A citation between two scientists should without any doubt have a different meaning when it is made after their co-authorship of many articles or long before they get to know each other. This enhancement might add even more "justice" to the ranking.

8. Acknowledgement

This work was partly supported by National Research Grant 2C06009

9. References

[1] Newman M. E. J. The Structure and Function of Complex Networks. SIAM Review, vol. 45, no. 2, pp. 167-256, 2003.

[2] Liu X., Bollen J., Nelson M. L., Van de Sompel H. Co-authorship Networks in the Digital Library Research Community. Information Processing and Management, vol. 41, no. 6, pp. 1462-1480, 2005.

[3] Brin S., Page L. The Anatomy of a Large-Scale Hypertextual Web Search Engine. Proceedings of the 7th World Wide Web Conference, pp. 107 – 117, 1998.

[4] Kleinberg J. Authoritative Sources in a Hyperlinked Environment. Journal of the ACM, vol. 46, no. 5, pp. 604-632, 1999.

[5] Fiala D., Tesar R., Jezek K., Rousselot F. Extracting Information from Web Content and Structure. Proceedings of the 9th International Conference on Information Systems Implementation and Modelling ISIM'06, PY´erov, Czech Republic, pp. 133-140, 2006.

[6] Page L., Brin S., Motwani R., Winograd T. The PageRank Citation Ranking: Bringing Order to the Web. Computer Science Department, Stanford University, California, USA, Technical Report 1999-66, Nov. 1999.

[7] Chakrabarti S. *Mining the Web: Analysis of Hypertext and Semi Structured Data*. Morgan Kaufmann Publishers, San Francisco, California, USA, 2002.

[8] Sidiropoulos A., Manolopoulos Y. A Citation-Based System to Assist Prize Awarding. SIGMOD Record, vol. 34, no. 4, pp. 54-60, 2005.

[9] Kleinberg J., Kumar R., Raghavan P., Rajagopalan S., Tomkins A. The web as a graph: Measurements, models and methods. Proceedings of the 5th Annual International Conference on Combinatorics and Computing, Tokyo, Japan, Lecture Notes in Computer Science, Springer, vol. 1627, pp. 1-17, 1999.

[10] Hirsch J. E. An index to quantify an individual's scientific research output. Proceedings of the National Academy of Sciences, vol. 102, no. 46, pp. 16569-16572, 2005.

[11] Bollen J., Rodriquez M. A., Van de Sompel H. *Journal status*. Scientometrics, vol. 69, no. 3, pp. 669-687, 2006.