A Citation-Based System to Assist Prize Awarding

Antonis Sidiropoulos
Data Engineering Lab, Department of Informatics,
Aristotle University, Thessaloniki, 54124 Greece
asidirop@csd.auth.gr

Yannis Manolopoulos
Data Engineering Lab, Department of Informatics,
Aristotle University, Thessaloniki, 54124 Greece
manolopo@csd.auth.gr

Abstract

Citation analysis is performed to evaluate the impact of scientific collections (journals and conferences), publications and scholar authors. In this paper we investigate alternative methods to provide a generalized approach to rank scientific publications. We use the SCEAS system [12] as a base platform to introduce new methods that can be used for ranking scientific publications. Moreover, we tune our approach along the reasoning of the prizes '*VLDB 10 Year Award*' and '*SIGMOD Test of Time Award*', which have been awarded in the course of the top two database conferences. Our approach can be used to objectively suggest the publications and the respective authors the are more likely to be awarded in the near future at these conferences.

1. Introduction

Citation analysis is performed to evaluate the impact of scientific collections (journals and conferences), publications and scholar authors during hiring, promotion, tenure and merit pay decisions. The first study in this field appeared in 1972 [4], however we meet many similar studies in the literature up to the present [10, 11].

In general, the ranking algorithms that are used in bibliometrics can be separated into two classes. We call the first one *collection-based ranking algorithms*. At this class of algorithms, a weighted citation graph is used, where the collections are the graph nodes, whereas the weighted edges represent the total number of the citations that point from one collection to another. The ISI Impact Factor belongs to this ranking class [3, 4, 5].

Our alternative approach, the SCEAS (Scientific Collection Evaluator with Advanced Scoring) method and system which has been presented in [12], also belongs in the same ranking class. SCEAS is a web-based digital library and its contents are imported from the Data Bases and Logic Programming (DBLP) website[1] of the University of Trier. The SCEAS system emerges as a useful tool for conducting several experiments on the DBLP data. Most of the results that are presented in this paper (as well as other results) are accessible through the SCEAS website[2].

We call the second class of ranking methods *publication-based ranking algorithms*. According to this approach, the nodes of the citation graph represent publications, whereas an edge from node x to node y represents a citation from paper x to paper y. The advantage of ranking at the publication level is that it is possible to evaluate more than one entity through a single computation: the paper itself, the collection where it belongs, and the author(s) as well. The last two evaluations can be made by an aggregate average of the first one. All the ranking algorithms that where initially proposed to rank web pages can be categorized in this class. In this respect, two of the most well known algorithms are the PageRank by Brin and Page [1] and HITS by Kleinberg's, which classifies the graph nodes as Hubs and Authorities [7, 8, 9].

The structure of this paper is as follows. In the next section, we briefly present the ranking methods for the *publication-based ranking* (PageRank and HITS) and we criticize their weakness when used in bibliometrics. We also investigate new alternative methods to overcome the vulnerabilities of the above algorithms, with respect to our specific goal. These new methods have been used within the SCEAS system, and thus we named our family of ranking algorithm as *SCEAS Rank*. Section 3 shows the algorithm performance with respect to the computation speed. Section 4 is divided in two parts, each part containing the rank results over the DBLP data for publications and authors respectively. In this section, we also evaluate the potential of the ranking methods by tuning and comparing their results to the well known awards '*VLDB 10 Year Award*' and '*SIGMOD Test of Time Award*', which have been given in the recent past in the two most prominent conferences of the database community.

2. Ranking Algorithms

In this section we introduce the well known PageRank and HITS algorithms and we investigate the reasons of their deficiency when they are used for publications ranking. We also present two new ranking methods.

2.1 HITS

HITS algorithm has been proposed as a method for ranking web pages that are retrieved when searching via a browser. It is based on two specific notions: *hubs* and *authorities*. In particular, the score of hubs and authorities is calculated by using the equations:

$$\begin{aligned} \vec{a'} &= A^T \vec{h} \\ \vec{h'} &= A * \vec{a} \end{aligned} \tag{1}$$

where $\vec{a}$ and $\vec{h}$ are vectors containing the scores of the publications as authorities and hubs respectively. A is the adjacency matrix of the citation graph.

Despite the popularity of HITS in the web context, HITS is not quite suitable in the field of bibliometrics. The reason is that a publication gets high authority score iff there are hubs pointing to it. However, this should not be the main metric in bibliometrics. Kleinberg proposed a model where authorities directly endorse other authorities in [7]. This concept it termed *Prestige* by Chakrabarti in [2]. However, after a closer look, it can be concluded that this approach seems to have problems when the graph does not include any cycles. In this case, the scores of the nodes converge to zero.

2.2 PageRank

PageRank is also popular in the web context. PageRank uses the following formula to calculate the score S_j of an object j (a page

[1] http://www.informatik.uni-trier.de/~ley/db/

[2] http://delab.csd.auth.gr/sceas

in the web context, or a publication in the context of bibliometrics):

$$S_j = (1-d) + d * \sum_{i \to j} \frac{S_i}{N_i} \quad (2)$$

where N_i is the number of citations from publication i, and d is a *damping factor*, usually set to 0.85.

PageRank associates high ranking to node i, if there is a big connected component C where some of its nodes point to i. The more and larger cycles C contains and i belongs to, the bigger score i will come up with. However, cycles do not usually exist in bibliometrics, but if they do so, they are usually self-citations. Thus, it is not fair for the self-citations to affect the score positively. On the other hand, removing the self-citations or the cycles will invalidate the graph and the results. Thus, PageRank could not be fair. This is verified by the graph instance with 12 nodes of Figure 1 and the results of Table 1, where we depict the ranking score for each node by applying all the ranking methods. Table 1 consists of five columns. Each column presents the ranking by the appropriate method and includes 2 sub-columns. Sub-column 1 is the *node id* and sub-column 2 is the *node score* rounded to 2 decimal digits. In this example, node 0 gets 4 citations, while nodes 10 and 6 get 3 citations each. However, the PageRank score of nodes 10 and 6 is about 3 times higher than the score of node 0. This happens because nodes 10 and 6 are parts of citation cycles.

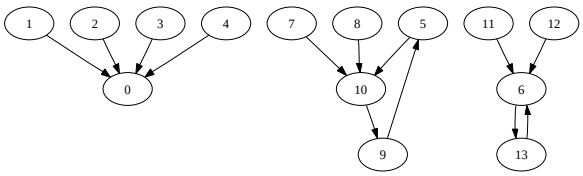

Figure 1. Example 1.

HITS_AU		PageRank		SCEAS1		SCEAS2		Citations	
0	1.00	6	1.92	0	1.47	0	0.34	0	4
1	0.00	13	1.78	6	1.43	6	0.32	6	3
2	0.00	10	1.66	10	1.36	10	0.32	10	3
3	0.00	9	1.56	13	0.90	13	0.25	5	1
4	0.00	5	1.48	9	0.87	9	0.25	9	1
5	0.00	0	0.66	5	0.69	5	0.23	13	1
6	0.00	1	0.15	1	0.00	1	0.15	1	0
7	0.00	2	0.15	2	0.00	2	0.15	2	0
8	0.00	3	0.15	3	0.00	3	0.15	3	0
9	0.00	4	0.15	4	0.00	4	0.15	4	0
10	0.00	7	0.15	7	0.00	7	0.15	7	0
11	0.00	8	0.15	8	0.00	8	0.15	8	0
12	0.00	11	0.15	11	0.00	11	0.15	11	0
13	0.00	12	0.15	12	0.00	12	0.15	12	0

Table 1. Rank tables for citation graph in Figure 1.

Secondly, PageRank is designed in a way (which is suitable for the web) that a page score is mostly affected by the scores of the web pages that point to it and less by the number of the incoming links (citations). A case like this is shown in Figure 2 and Table 2, where node 0 gets higher score than node 1, although node 1 gets 6 citations.

In addition to the above cases, any change of node j score affects the score of node i even if the path connecting them is very long. Neither is required in bibliometrics. In other words, in cases where direct citations exist, it is necessary for the impact to be (a) large and (b) much less when the distance between the two nodes gets larger. This is depicted in Figure 3. Adding a link to node 6 in Figure 4 results in a 7.14% increase of node 5 score and 6.82% of node 4 score, which are quite distant. This is yet another case where PageRank has not been proved to have good behavior in bibliometrics.

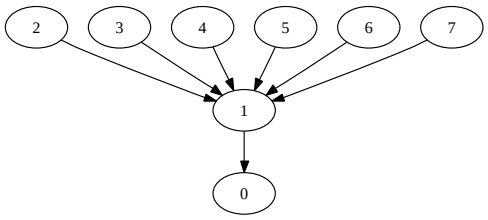

Figure 2. Example 2.

HITS_AU		PageRank		SCEAS1		SCEAS2		Citations	
1	1.00	0	0.93	1	2.21	1	0.43	1	7
0	0.00	1	0.92	0	1.18	0	0.29	0	1
2	0.00	2	0.15	2	0.00	2	0.15	2	0
3	0.00	3	0.15	3	0.00	3	0.15	3	0
4	0.00	4	0.15	4	0.00	4	0.15	4	0
5	0.00	5	0.15	5	0.00	5	0.15	5	0
6	0.00	6	0.15	6	0.00	6	0.15	6	0
7	0.00	7	0.15	7	0.00	7	0.15	7	0

Table 2. Rank tables for citation graph in Figure 2.

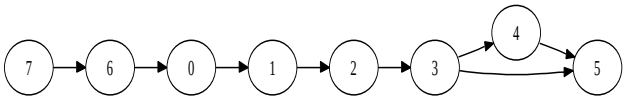

Figure 3. Example 3.

HITS_AU		PageRank		SCEAS1		SCEAS2		Citations	
5	0.85	5	0.77	5	0.76	5	0.24	5	2
4	0.53	3	0.62	3	0.58	3	0.22	0	1
0	0.00	2	0.56	2	0.57	2	0.22	1	1
1	0.00	1	0.48	1	0.55	1	0.22	2	1
2	0.00	4	0.41	0	0.50	0	0.21	3	1
3	0.00	0	0.39	6	0.37	6	0.20	4	1
6	0.00	6	0.28	4	0.29	4	0.18	6	1
7	0.00	7	0.15	7	0.00	7	0.15	7	0

Table 3. Rank tables for citation graph in Figure 3.

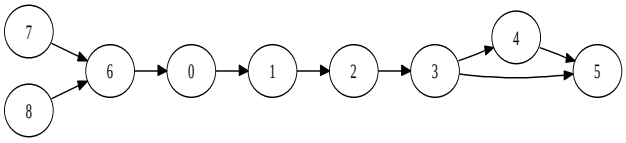

Figure 4. Figure 4.

HITS_AU		PageRank		SCEAS1		SCEAS2		Citations	
5	0.85	5	0.82	5	0.77	6	0.24	5	2
4	0.53	3	0.69	6	0.74	5	0.24	6	2
6	0.00	2	0.63	0	0.64	0	0.23	0	1
0	0.00	1	0.57	1	0.60	1	0.22	1	1
1	0.00	0	0.49	2	0.59	2	0.22	2	1
2	0.00	4	0.44	3	0.58	3	0.22	3	1
3	0.00	6	0.41	4	0.29	4	0.18	4	1
7	0.00	7	0.15	7	0.00	7	0.15	7	0
8	0.00	8	0.15	8	0.00	8	0.15	8	0

Table 4. Rank tables for citation graph in Figure 4.

2.3 SCEAS Rank

In this section, under the generic name SCEAS Rank, we propose two new ranking techniques, named SCEAS1 and SCEAS2. All the requirements mentioned in the previous section can be embedded

in a ranking score, S_j for node j, as follows:

$$S_j = \sum_{i \to j} \frac{S_i + b}{N_i} a^{-1} \quad (a \geq 1, b > 0) \tag{3}$$

where N_i denotes the number of outgoing citations of node i, b is the *direct citation enforcement factor* and factor a denotes the speed in which an indirect citation enforcement converges to zero. The above formula depicts the fact that a change of node i score, affects the score of node j that is x nodes away, by the factor a^{-x}. For our tests we have selected a to be equal to e. Therefore, the involvement of node i to the score of node j is less as the distance x between them gets longer. Actually, it converges to zero for $x > 7$.

In Equation 3 we have used the factor b since citations coming from zero scored nodes should also contribute in the score of their citing publications. The value of b may vary, according to the level that the number of the incoming citations should affect the calculated score. In any case, it must be greater than 0, otherwise the scores will converge to zero. In our tests we have used $b = 1$ as the number of the incoming citations should count significantly.

As shown in the tables above for all the examples, the results of SCEAS1 are satisfactory enough. In Table 1, node 0 is ranked first as it gets 4 citations. Nodes 6 and 10 are ranked second and third, respectively. Node 6, like node 10, takes part in citation cycles. As a result their scores are affected by themselves. Fortunately, the score increment that they gain due to cycles is not that high to advance them to the first position. Thus, in this example, SCEAS1 gives better results than PageRank. In Table 2, SCEAS1 puts node 1 in the first place in contrast to PageRank that ranks node 0 in the first place. In the examples of Table 3 and 4, after node 6 is cited by the just added node 8, SCEAS1 places node 6 in the second place. Also, the node 5 score increment is only 1% relatively to its previous one. On the other hand the change of the score of node 6 is 100% (doubled score), since the number of citations that point to it has been doubled. This small example shows clearly that when new nodes and citations are added to the citation graph, the score of the nodes that are far away from the new nodes are not affected significantly. This means that a very fast incremental computation is feasible.

In summary, the advantages of Equation 3 over PageRank and HITS are:

- The score of a node is mainly affected by the number of incoming citations.
- Computation and convergence is very fast (the results are shown in Section 3).
- The score of a node is less affected by the score of distant nodes. This has also the effect that when new nodes and citations are added, the new computation of the score can be incremental by using the previous score vector as initial vector. The new node actually influences the scores of nodes at most 7 nodes away. That's because practically e^{-7} is almost 0 compared with the scores of nodes. So, the incremental computation is very fast, since the new node influences only a small part of the overall graph.

A dumping factor d may be added to Equation 3 without any major effect in the rank results. This would lead to the following equation, which is a generalized formula of SCEAS1 and PageRank:

$$S_j = (1 - d) + d * \sum_{i \to j} \frac{S_i + b}{N_i} a^{-1} \quad (a \geq 1) \tag{4}$$

For $d = 1$ Equation 4 is equivalent to Equation 3. For $b = 0$ and $a = 1$ Equation 4 is equivalent to PageRank. PageRank uses the value of $d = 0.85$, since this value balances the precision and the convergence speed. A value d closer to 1, means better precision for the scores. Also, a value $d = 1$ should lead PageRank to converge to zero. Therefore, a value $d < 1$ is necessary for PageRank. In our generalized formula, it is safe to use any value for d ($0 < d \leq 1$) if $b > 0$. Also, the convergence speed is mainly affected by the factor a rather than d. In our case, it is safe to use a greater factor d, such as 0.99.

In the examples of Tables 1, 2, 3 and 4 the column for SCEAS1 assumes that $d = 1, b = 1$ and $a = e$, whereas SCEAS2 assumes that $d = 0.85, b = 0$ and $a = e$. A rank with $d = 0.85, b = 1$ and $a = e$ is not included in the tests, since this gives equivalent results to SCEAS1 for the dataset of DBLP digital library.

3. SCEAS Rank Speed

According to the definition of SCEAS Rank, it is obvious that we come up with a very fast convergence by using $b = 1$ and $a = e$. In Figure 5, the x-axis represents the number of the iterations that are needed by each algorithm to compute the ranks for the DBLP digital library. Axis y shows the value of

$$\delta = ||\vec{x}_l - \vec{x}_{l-1}||_1 \tag{5}$$

where $\vec{x}_l$ is the vector with the scores $\{S_1, S_2, ...S_V\}$ after l iterations. The termination condition for each algorithm is: $\delta < \epsilon$, where ϵ is a very small number. Actually, as described in [6], this number could not be predefined for PageRank since it depends on the citation graph. It is obvious that each algorithm needs a different value of ϵ as a termination condition. Regardless of this, the plot shows that the lines of SCEAS Rank are much steeper than the other algorithm lines. This means that it converges faster than the other methods no matter what the actual values of δ and ϵ are.

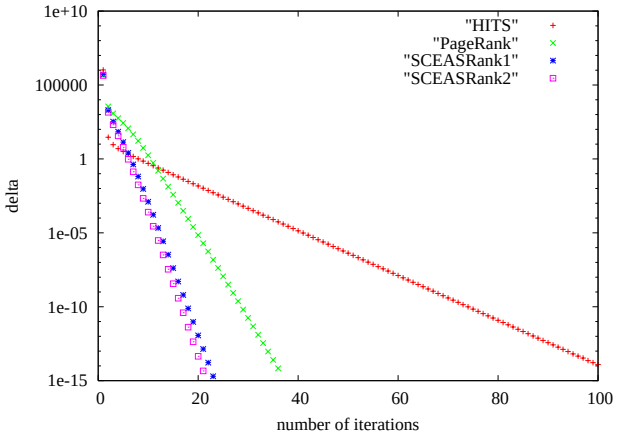

Figure 5. Convergence speeds for DBLP.

Note here that HITS is even slower than what is illustrated in Figure 5. Each iteration of HITS requires about twice the time needed for one iteration of PageRank or SCEAS. This happens because there are two vector multiplications plus a normalization step during each iteration of HITS. The line of HITS should be shifted upwards (as multiplied by 2 at least) to be comparable to SCEAS and PageRank. Conclusively, we can say that SCEAS needs about half the time needed by PageRank and about 1/10 of the time needed by HITS with respect to the DBLP digital library.

4. Results

In this section we present some results on publication and author ranking using data form the DBLP digital Library and our SCEAS

system. Here, it is worth noticing that the DBLP data set is incomplete. The graph out-degrees (citations) are available only for a small part of the total number of the publications contained in the database.

In particular, SCEAS uses the DBLP data included in the inproceedings (370790) and in the articles (217950) collections, which makes a total of 588865 publications as of when our experimentation took place (latest timestamp in DBLP records is 24th Jan 2005). Only 8183 of them (i.e., 1.3%) have their citations actually stored (publications with out-degree). These papers are spanning the time window 1970-2000. The total number of publications during this period is 376912, which means that the citations input is still small. In addition, 18273 publications have in-degree, and thus actually these publications are being ranked.

The total number of citations is 167999 (i.e., the 8183 publications have an average out-degree equal to 20.5). Also, 100210 (i.e., 59.65%) of the above citations point to a publication into the DBLP digital library. The remaining 67789 citations (i.e., 40.35%) point to papers outside the DBLP digital library, and thus they are ignored from further consideration.

Journal	#citations	#cit. in DBLP		#papers	period
ACM Computing Surveys	4471	2733	61%	68	1976-1998
ACM SIGMOD Digital Review	196	185	94%	146	1999-2001
ACM Trans. Database Systems	13313	8011	60%	497	1976-1999
Communications ACM	501	302	60%	31	1970-1999
IEEE Data Eng. Bullettin	2545	1450	57%	178	1991-1999
IEEE Trans. Knowl. Data Eng.	19355	10552	55%	667	1989-1998
SIGMOD Record	1817	1062	58%	99	1981-2000
VLDB Journal	5112	3380	66%	132	1992-2000
journal Summary	47310	27675	58%	1818	1970-2001

Table 5. Journals with citations (out-degree) in DBLP.

Conference	#citations	#cit. in DBLP		#papers	period
ACM SIGMOD	19129	11763	61%	1083	1975-2000
ADBIS	3803	1788	47%	233	1994-1999
CIKM	980	601	61%	55	1995-1995
CoopIS	538	238	44%	30	2000-2000
DASFAA	5042	3131	62%	307	1989-1999
DBPL	3506	2303	66%	148	1987-1997
Digital Libraries	559	170	30%	33	1997-1997
DOLAP	360	223	62%	27	1998-1999
EDBT	4589	2953	64%	244	1988-2000
ER	13267	7114	54%	664	1979-1999
ICDE	18552	11799	64%	1064	1984-1999
ICDT	5239	3593	69%	235	1986-2001
MFDBS	1319	844	64%	68	1987-1991
PDIS	463	307	66%	31	1994-1994
PODS	13081	8580	66%	606	1982-2000
POPL	73	47	64%	4	1979-1982
RIDE	219	113	52%	13	2001-2001
SSDBM	3902	1792	46%	230	1981-1999
VLDB	25887	15059	58%	1277	1975-2000
WIDM	181	117	65%	13	1999-1999
Conferences Summary	120689	72535	60%	6365	1975-2001

Table 6. Conferences with citations (out-degree) in DBLP.

However, despite the above remarks, there still remains a very good sample for our ranking purposes as there is no lack of data with respect to the highly competitive conferences under consideration. Tables 5 and 6 show in a detailed manner the conferences and journals respectively, which have their citations stored, how many citations exist for each one of them, how many of these citations point into the DBLP digital library, how many papers each journal or conference comprises of and, finally, the respective time period. In any case, it is very difficult to collect all the citations to a specific publication. As shown in Figure 6, the number of citations to papers published during the period 1986-1994 is quite high. This gives a good sample for the specific period of time. Unfortunately, the number of citations decreases for years after 1995. Therefore, we will not present the rank results for papers published after 1995, since the sample may be considered inadequate.

Year	Title	P	S	H	C
1986	Object and File Management in the EXODUS Extensible Database System. Michael Carey, David DeWitt, Joel Richardson, Eugene Shekita	2	3	1	2
1987	The R+-Tree: A Dynamic Index for Multi-Dimensional Objects. Timos Sellis, Nick Roussopoulos, Christos Faloutsos	1	1	1	1
1988	Disk Shadowing. Dina Bitton, Jim Gray	1	1	5	2
1989	ARIES/NT: A Recovery Method Based on Write-Ahead Logging for Nested Transactions. Kurt Rothermel, C. Mohan	6	7	14	6
1990	Deriving Production Rules for Constraint Maintainance. Stefano Ceri, Jennifer Widom	1	1	3	1
1991	A Transactional Model for Long-Running Activities. Umeshwar Dayal, Meichun Hsu, Rivka Ladin	2	2	24	4
1992	Querying in Highly Mobile Distributed Environments. Tomasz Imielinski, B.R. Badrinath	3	4	43	12
1993	Universality of Serial Histograms. Yannis Ioannidis	6	7	8	5
1994	Fast Algorithms for Mining Association Rules in Large Databases. Rakesh Agrawal, Ramakrishnan Srikant	1	1	4	1
		23	27	103	35

Table 7. Rank positions of publications awarded with the VLDB 10 Year Award.

Our results are separated in two parts. In the first part we apply the algorithms mentioned earlier to rank/evaluate publications. The second part is dedicated to rank/evaluate authors.

4.1 Ranking Publications

It is a very tough task to evaluate a ranking algorithm for scientific publications, since it is rather subjective which ranking algorithm behavior is better. As a criterion to verify that our ranking results are appropriate, we used two well known awards for publications: the 'VLDB 10 Year Award' and the 'SIGMOD Test of Time Award'. We accept that if any algorithm gives high rank positions to the awarded publications, then this algorithm can be safely used for the task of evaluating publications.

We compare four rank methods: PageRank, SCEAS Rank, HITS Rank (Authorities) and finally the number of Citations. Particular SCEAS Rank variations are not presented here, since they both produce similar results for the data of VLDB and SIGMOD conferences. The tested scenario is the following:

1. Perform all the rank algorithms on the DBLP citation graph.
2. Get only the papers which are inproceedings of VLDB and SIGMOD conferences.
3. Organize and sort the rank tables grouped by conference and year.
4. Check the position of the awarded papers in the above rank tables.

In Tables 7 and 8 we show the awarded publications and their rank position for all of the ranking methods. In these tables, column P stands for PageRank, column S for SCEAS Rank, column H for HITS Rank (Authorities) and column C for the plain Citation counter. For example the paper entitled 'Fast Algorithms for Mining Association Rules in Large DBs, 1994' (Table 7) is ranked first by PageRank, first by SCEAS, fourth by HITS (as an authority) and first by plain Citation counter. Notice again that this way we do not evaluate the awarding committees but the rank methods.

In these detailed tables, we can see that the awarded papers are generally highly ranked. Apparently some deviations and exceptions do exist. These exceptions may exist for two reasons:

- Our citations sample may be not enough: e.g. an awarded publication may get a lot of citations from scientific domains that are not included in the DBLP digital library.

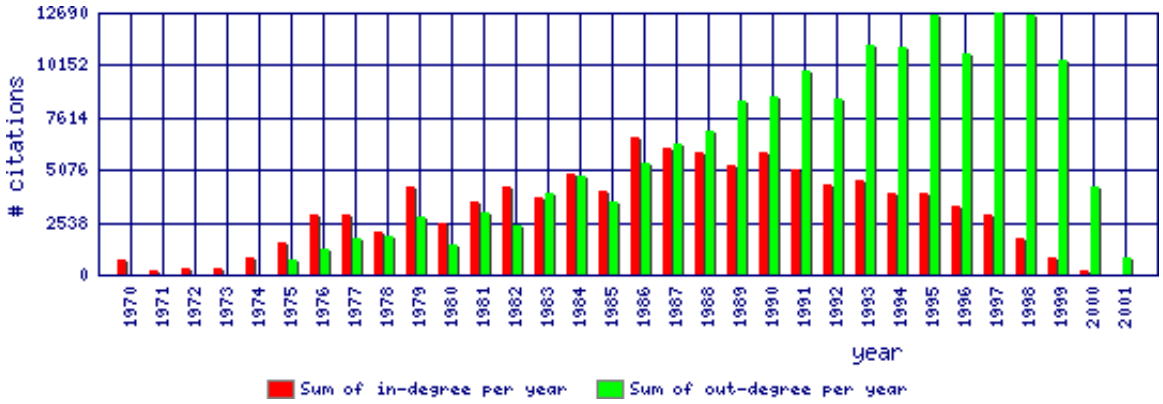

Figure 6. Citations distribution over time (all conferences and journals).

Year	Title	P	S	A	c
1988	A Case for Redundant Arrays of Inexpensive Disks (RAID). David Patterson, Garth Gibson, Randy Katz	1	1	8	2
1989	F-Logic: A Higher-Order language for Reasoning about Objects, Inheritance, and Scheme. Michael Kifer, Georg Lausen	6	4	5	5
1990	Encapsulation of Parallelism in the Volcano Query Processing System. Goetz Graefe	10	9	5	11
1990	Set-Oriented Production Rules in Relational Database Systems. Jennifer Widom, Sheldon Finkelstein	3	3	4	3
1992	Extensible/Rule Based Query Rewrite Optimization in Starburst. Hamid Pirahesh, Joseph Hellerstein, Waqar Hasan	1	1	2	2
1992	Querying Object-Oriented Databases. Michael Kifer, Won Kim, Yehoshua Sagiv	2	3	1	2
1993	Mining Association Rules between Sets of Items in Large Databases. Rakesh Agrawal, Tomasz Imielinski, Arun Swami	1	1	6	1
1994	From Structured Documents to Novel Query Facilities. Vassilis Christophides, Serge Abiteboul, Sophie Cluet, Michel Scholl	2	2	7	2
1994	Shoring Up Persistent Applications. Michael Carey, David DeWitt, Michael Franklin, Nancy Hall, Mark McAuliffe, Jeffrey Naughton, Daniel Schuh, Marvin Solomon, C.K. Tan, Odysseas Tsatalos, Seth White, Michael Zwilling	1	1	1	1
		27	25	39	29

Table 8. Rank positions of publications awarded with the SIGMOD Test of Time Award.

- The awards may be by definition subjective: e.g. an awards committee decision may be based on objective factors (such as citations to papers) but also may combine other measures and indicators of impact.

To get an overall view of the performance of the rank methods, we sum the positions of the awarded publications in the last row. The smaller this sum, the better performance of the ranking method. In both tables, we remark that the HITS Ranking (Authorities) is by far the worst method. This verifies our remarks explained in Section 2. Among the other three methods, the plain citation count is the weakest method, but it still remains a quite good evaluation method. Finally, we see that PageRank and SCEAS Rank are very close to each other and alternate at the winning position in the two Tables 7 and 8. In the sequel, we will ignore the HITS Rank (Authorities) in our tests, since it is not suitable for our evaluation purposes.

Since PageRank, SCEAS Rank and Citation count ended up with about similar results, we will try to produce a single rank table by averaging their results along the reasoning of [11]. In simple words, we will compute all the three rankings and assign to each paper a number of points (5 to 1) depending on its position in the specific rank table. For example, in each table the first paper gets 5 points, the second one gets 4 points, and so on. Thus, if a paper is ranked first in all three rankings, then it will get 15 points. Therefore, we repeat steps 3 and 4 of the previous scenario to produce the new rank tables. This computation is shown in Tables 9 and 10 for the awarded publications. It can be easily seen that the majority of the awarded publications are ranked in the top 3 positions of these new rank tables. Also, after exhaustive experiments we concluded that the sum of the positions (column Pos in both tables) is smaller by using this averaged approach in comparison to using any stand alone rank method. In particular, we remark that in the SIGMOD case (Table 10) a smaller sum of positions is achieved (i.e. 24) in comparison to the VLDB case (Table 9) where the sum is greater (i.e. 27) largely due to the 1989/1999 outlier.

Year	Title	Pos	Points
1986	Object and File Management in the EXODUS Extensible Database System. Michael Carey, David DeWitt, Joel Richardson, Eugene Shekita	2	11
1987	The R+-Tree: A Dynamic Index for Multi-Dimensional Objects. Timos Sellis, Nick Roussopoulos, Christos Faloutsos	1	15
1988	Disk Shadowing. Dina Bitton, Jim Gray	1	14
1989	ARIES/NT: A Recovery Method Based on Write-Ahead Logging for Nested Transactions. Kurt Rothermel, C. Mohan	9	0
1990	Deriving Production Rules for Constraint Maintainance. Stefano Ceri, Jennifer Widom	1	15
1991	A Transactional Model for Long-Running Activities. Umeshwar Dayal, Meichun Hsu, Rivka Ladin	2	10
1992	Querying in Highly Mobile Distributed Environments. Tomasz Imielinski, B.R. Badrinath	4	5
1993	Universality of Serial Histograms. Yannis Ioannidis	6	1
1994	Fast Algorithms for Mining Association Rules in Large Databases. Rakesh Agrawal, Ramakrishnan Srikant	1	15

Table 9. Sum of rank positions of publications awarded with the VLDB 10 Year Award.

4.2 Ranking Authors

We may rely on our method of computing scores for publications and compute scores for authors as well. An approach could be to compute the average score of all their publications. This is not a trivial task. For example, author A has 200 publications with only 40 ones being 'first class'. Assume that these high quality publications have a score of 10 points each, whereas the remaining ones have a score of 1 point. Author B has in total 20 publications, with 10 publications of them being 'first class'. It is reasonable to consider that author A should be ranked higher than author B for his scientific contribution, because A has 4 times the number of first class publications than author B. However, if we just compute the average of all publication scores, then authors A and B would have 2.8 and 5.5 points respectively. Therefore, it is not fair to take into account all the publications a person has authored. In addition, it is not fair to take into account different number of publications

Author Name	Rank Position by average of top-x publications of each author															
	1	2	3	4	5	6	7	8	9	10	12	15	20	25	30	40
C. Mohan	112	122	108	93	83	81	71	69	65	62	54	49	42	34	32	26
David DeWitt	33	26	21	16	15	14	13	13	12	12	12	7	7	5	5	3
David Maier	43	29	29	27	21	19	17	17	16	16	15	13	12	11	12	8
Donald Chamberlin	9	8	4	4	4	4	4	4	4	4	4	4	5	7	8	
Hector Garcia-Molina	88	53	46	41	38	34	33	30	29	27	24	22	22	19	18	12
Jim Gray	3	4	3	2	2	2	2	2	2	2	2	2	2	2	1	1
Michael Stonebraker	24	11	10	9	9	8	8	7	5	5	5	5	4	4	3	2
Patricia Selinger	7	21	23	24	24	23	25	26	25	26	27	32	33	32	34	
Philip Bernstein	63	36	28	21	18	15	14	14	11	11	11	8	8	8	7	5
Rakesh Agrawal	53	37	35	31	27	24	22	22	20	20	19	17	14	12	11	7
Ronald Fagin	49	34	30	26	22	20	19	18	18	18	17	15	17	14	15	13
Rudolf Bayer	25	24	26	28	29	31	31	33	34	38	38	35	37	37	41	34
Serge Abiteboul	165	91	66	54	48	44	43	41	40	36	31	26	25	22	21	15
Lowest ranking point	165	122	108	93	83	81	71	69	65	62	54	49	42	37	41	34
Sum of ranking points	674	496	429	376	340	319	302	296	281	277	259	235	228	207	208	

Table 11. SCEAS Rank average scores for authors.

Year	Title	Pos	Points
1988	A Case for Redundant Arrays of Inexpensive Disks (RAID). David Patterson, Garth Gibson, Randy Katz	1	14
1989	F-Logic: A Higher-Order language for Reasoning about Objects, Inheritance, and Scheme. Michael Kifer, Georg Lausen	5	3
1990	Encapsulation of Parallelism in the Volcano Query Processing System. Goetz Graefe	7	0
1990	Set-Oriented Production Rules in Relational Database Systems. Jennifer Widom, Sheldon Finkelstein	3	9
1992	Extensible/Rule Based Query Rewrite Optimization in Starburst. Hamid Pirahesh, Joseph Hellerstein, Waqar Hasan	1	14
1992	Querying Object-Oriented Databases. Michael Kifer, Won Kim, Yehoshua Sagiv	3	11
1993	Mining Association Rules between Sets of Items in Large Databases. Rakesh Agrawal, Tomasz Imielinski, Arun Swami	1	15
1994	From Structured Documents to Novel Query Facilities. Vassilis Christophides, Serge Abiteboul, Sophie Cluet, Michel Scholl	2	12
1994	Shoring Up Persistent Applications. Michael Carey, David DeWitt, Michael Franklin, Nancy Hall, Mark McAuliffe, Jeffrey Naughton, Daniel Schuh, Marvin Solomon, Tan, Odysseas Tsatalos, Seth White, Michael Zwilling	1	15

Table 10. Sum of rank positions of publications awarded with the SIGMOD Test of Time Award.

for each author (e.g. 40 publications for A and 10 for B). In our approach, we take into account the same number of publications for all authors so that the score results could be comparable.

Therefore, our problem now is to choose the number of publications of each author that should be considered in the ranking. To determine this number we performed the following experiment. We computed the average score for each author by using his top 1-10, ..., 40 publications. Thus, we produced 16 rankings for every rank method. As a testbed we used the authors that were awarded the '*SIGMOD Edgar F. Codd Innovations Award*'. The higher these authors were ranked, the better the evaluation was considered. In Table 11 we present the results that were produced by SCEAS Rank. For brevity, we present only the awarded authors and their position for each selected number of 'top' publications. For example, Hector-Garcia Molina is ranked 88th in the ranking, if the score is produced by the average of 1-best publication of each author. He is ranked 53rd in the ranking, if the score is produced by the average of 2-best publications, and so on.

The last two rows of Table 11 show the rank position of the awarded author that ranked last ('lowest ranking point') and the 'sum of ranking positions' of all the awarded authors. These two numbers are our metrics for comparing the rankings. The lower these numbers are, the best ranking is performed. The lowest ranking point is low indeed when computing the average for the 1-4 best publications of each author. That is due to the fact that the co-authors of a 'high class' publication take advantage and climb up the ranking results. Therefore, by increasing the number of the selected 'best' publications, the awarded authors move towards the top of the rank table. This trend holds until the number of the selected publications becomes 25. The same remark holds when we consider the notion of 'sum of ranking positions'. Also note that in column 40 we miss 2 of the awarded authors, since they have less than 40 publications in the DBLP digital library. For brevity, we have not included the respective tables for the other rank methods, since the results are similar and the smallest ranking point is the same. Thus, after this experiment we decided to rank the authors by averaging their 25 best publications. A final remark with respect to this table is the following. Quite interesting and easily explained is the fact that there are several authors whose ranking position gets higher with increasing number of 'best' publications (with S. Abiteboul advancing the most), whereas the opposite holds for a few other authors (e.g. P. Selinger due to her work on System R and R. Bayer due to his work on B-trees and related structures). Jim Gray steadily holds top positions.

In Table 12 we compare the various rank methods. In this table we present the rank positions of the awarded authors for each Rank method by taking into account the average score of the 25 best publications of each author. In this table column P stands for PageRank, H for HITS Authorities, C for Citation count and S1 and S2 for SCEAS1 and SCEAS2 respactively. Column S1 of this table, is obviously equal to column '25' of Table 11. Similarly to the previous table, we compute the 'Lowest Ranking Point' and 'Sum of ranking points'. In this table we can see that HITS authorities is by far the worst method again, since the 'sum of ranking points' and the 'lowest ranking point' are about 2 and 3 times greater than the respective numbers computed for the other methods. Plain citation count gives an acceptable 'Sum of ranking points', but it fails in the 'lowest ranking point' metric. Finally, SCEAS1 and SCEAS2 are the best methods with respect to the 'lowest ranking point' metric and competitive to PageRank based on the 'Sum of ranking points' metric.

5. Conclusion

In this report we proposed and experimentally examined SCEAS Rank, a new alternative method for scientific publications evaluation, besides the known algorithms of PageRank and HITS. We also

	P	S1	H	S2	C
C. Mohan	41	34	62	33	29
David DeWitt	8	5	4	5	1
David Maier	13	11	8	11	9
Donald Chamberlin	5	7	11	7	14
Hector Garcia-Molina	20	19	115	19	21
Jim Gray	2	2	9	2	3
Michael Stonebraker	4	4	2	4	2
Patricia Selinger	28	32	24	34	38
Philip Bernstein	6	8	6	8	6
Rakesh Agrawal	16	12	82	12	15
Ronald Fagin	10	14	17	14	17
Rudolf Bayer	24	37	117	40	73
Serge Abiteboul	23	22	16	21	13
Lowest Ranking point	41	37	117	40	73
Sum of ranking points	200	207	473	210	241

Table 12. Rank position of awarded authors by average of 25 best publications.

presented SCEAS Rank tuned variations. However, detailed algorithm descriptions and performance tuning appears in [13]. We also evaluated the above method by using the DBLP digital library as a training set and the awarded publications of '*VLDB 10 Year Award*' and '*SIGMOD Test of Time Award*' as an evaluation set for the publications rank method. Additionally, we presented author ranking based on the publication rank results and we used the '*SIGMOD Edgar F. Codd Innovations Award*' as an evaluation set. In both cases the performance of SCEAS Rank was in general better than the other methods. This might be helpful to short-list candidate authors for awarding during the next few years.

References

[1] S. Brin and L. Page. The Anatomy of a Large-scale Hypertextual Web Search Engine. In *Proceedings 7th WWW Conference*, pages 107–117, 1998.

[2] S. Chakrabarti. *Mining the Web: Discovering Knowledge from Hypertext Data*, chapter 7.1, pages 205–206. Morgan Kaufmann Publishers, 2003.

[3] E. Garfield. Science citation index. http://www.isinet.com/isi/.

[4] E. Garfield. Citation Analysis as a Tool in Journal Evaluation. *Essays of an Information Scientist*, 1:527–544, 1972.

[5] E. Garfield. The Impact Factor, 1994. http://www.isinet.com/isi/hot/essays/journalcitationreports/7.html.

[6] S. D. Kamvar and T. H. Haveliwala. The condition number of the pagerank problem. Technical report, Stanford University, 2003.

[7] J. Kleinberg. Authoritative Sources in a Hyperlinked Environment. *Journal of the ACM*, 46(5):604–632, 1999.

[8] J. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. The Web as a Graph: Measurements, Models, and Methods. In *Proceedings 5th COCOON Conference*, pages 1–17, 1999.

[9] J. Kleinjnen and W. V. Groenendaal. Measuring the Quality of Publications: New Methodology and Case Study. *Information Processing and Management*, 36(4):551–570, 2000.

[10] N. Mylonopoulos and V. Theoharakis. Global Perceptions of IS Journals. *Communications of the ACM*, 44(9):29–33, 2001.

[11] R. Rainer and M. Miller. Examining Differences across Journal Rankings. *Communications of the ACM*, 48(2):91–94, 2005.

[12] A. Sidiropoulos and Y. Manolopoulos. A New Perspective to Automatically Rank Scientific Conferences Using Digital Libraries. *Information Processing and Management*, 41(2):289–312, 2005.

[13] A. Sidiropoulos and Y. Manolopoulos. Ranking Algorithms for Publications and Authors. Technical report, Aristotle University, 2005. under submission.