Proceedings of the BIS 2007 Workshop on Social Aspects of the Web (SAW), pp. 27-43

AGE DECAYING H-INDEX FOR SOCIAL NETWORKS OF CITATIONS

Dimitrios Katsaros[1] Antonis Sidiropoulos[2] Yiannis Manolopoulos[2]
Dept. of Computer & Communication Engineering[1] *University of Thessaly* *Volos, Greece*
[2]*Department of Informatics* *Aristotle University* *Thessaloniki, Greece*

ABSTRACT

Mining the link structure of social networks is a relatively old topic dating back to 1950, but the field has recently gained tremendous attention, due to the compilation of huge amounts of data and the development of sophisticated computerized methods to process these data. Among the many facets of these mining tasks in various types of social networks, the mining task of ranking researchers or conferences/journals in citation networks is of particular importance and challenge due to its uses in many practical situations, e.g., faculty promotion, funding. The development of a fair metric to perform the ranking is a very hard task and a lot of methods have appeared in the literature. Recently, the method that has stimulated the interest of the research community is the h-index method. The h-index metric corrects a lot of the older methods inefficiencies. But, the h-index presents a major disadvantage: it keeps increasing even if the researcher has stopped publishing good papers or if his present work is not that significant. In this paper, we present a generalization of the basic h-index, termed age decaying h-index to cope with this situation. The new index is able to identify the researchers and publication forums that are perceived as very significant for the latest years. We present an extensive experimental evaluation of the proposed novel citation index over the DBLP bibliographic database. The evaluation proves the robustness and virtues of the new index and reveals some very interesting results concerning the performance of scientists, the emergence of research topics, and the preference of scientists to publish in new forums.

KEYWORDS

Hirsch index, h-index, ranking, bibliometrics, link mining, social networks.

1. INTRODUCTION

The academic community works on the evaluation of the scientists since the decade of 1970 (Garfield, 1972). Lately, due to the huge development of the digital libraries, the research efforts on the topic have multiplied, and a large number of researchers is involved in this topic (Barnes, 2005; Bernstein et al., 2005; Bharati and Tarasewich, 2002; Katerattanakul et al., 2003; Kelly Rainer and Miller, 2005; Lowry et al., 2004; Mylonopoulos and Theoharakis, 2001; Nascimento et al., 2003; Nerur et al., 2005; Rahm and Thor, 2005; Schwartz and Russo, 2004; Sidiropoulos and Manolopoulos, 2005*a*).

So far there are two major methods for the scientists' evaluation. The first method is based on the polling. A group of people has to be interviewed for their evaluation. The bigger the sample of people is, the better the evaluation that will be returned is. These works are very interesting, because they perform a ranking according to readers' (and authors') perception, but they suffer from the fact of being basically "manual" and usually biased, and not highly computerized and objective. The second method is based on the social network theory and is conducted through the citation analysis. The evaluation of the scientific work is performed by defining an objective function that calculates some "score" for the "objects" under evaluation, analyzing the social network formed by the citations among the published articles. Defining a quality and representative metric is not an easy task, since it should account for the productivity of a scientist and the impact of all of his/her work (analogously for journals/conferences). Most of the existing methods up-to-date are based on some form of (arithmetics upon) the total number of authored papers, the average number of authored papers per year, the total number of citations, the average number of citations per paper, the average number of citations per year, etc.

Finally, characteristic works implementing the hybrid approach of combining the experts' judge and citation analysis are described in (Kelly Rainer and Miller, 2005; Sidiropoulos and Manolopoulos, 2006). Their rankings are realized by taking some averages upon the results obtained from the citation analysis and experts' opinion, thus implementing a post-processing step of the two major approaches.

1.1 H-index and variations

Although, there is no clear winner among citation analysis and experts' assessment, the former is usually the preferred method, because it can be performed in a fully automated and computerized manner and it is able to exploit the wealth of citation information available in digital libraries.

All the metrics used so far in citation analysis present one or more drawbacks. These drawbacks have been presented by Hirsch (2005) and Sidiropoulos et al. (2007).

To collectively overcome all these disadvantages of the present metrics, during 2005 J. E. Hirsch proposed the pioneering *h-index* (Ball, 2005; Hirsch, 2005), defined as follows[1]:

Definition 1 *A researcher has h-index h if h of his/her N_p articles have received at least h citations each, and the rest $(N_p - h)$ articles have received no more than h citations.*

This metric calculates how broad the research work of a scientist is. The *h-index* accounts for both productivity and impact. For some researcher, to have large *h-index*, s/he must have a lot of "good" articles.

The *h-index* acts as a lower bound on the real number of citations for a scientist. Think that the quantity h will always be smaller than or equal to the number N_p of the articles of a researcher; it holds that $h^2 \leq N_{c,tot}$, where $N_{c,tot}$ is the total number of citations that the researcher has received. Apparently, the equality holds when all the articles, which contribute to *h-index* have received exactly h citations each, which is quite improbable. Therefore, in the

[1]Notice that the economics literature defines the H-index (the Herfindahl-Hirschman index), which is a way of measuring the concentration of market share held by particular suppliers in a market. The H index is the sum of squares of the percentages of the market shares held by the firms in a market. If there is a monopoly, i.e., one firm with all sales, the H index is 10000. If there is perfect competition, with an infinite number of firms with near-zero market share each, the H index is approximately zero. Other industry structures will have H indices between zero and 10000.

usual case it will hold that $h^2 < N_{c,tot}$. To bridge this gap, J. E. Hirsch defined the index a as follows:

Definition 2 *A scientist has a-index a if the following equation holds (Hirsch, 2005):*

$$N_{c,tot} = ah^2. \tag{1}$$

The *a-index* can be used as a second metric-index for the ranking of scientists. It describes the "magnitude" of each scientist's "hits". A large a implies that some article(s) have received a fairly large number of citations compared to the rest of its articles.

The introduction of the *h-index* was a major breakthrough in citation analysis. Though several aspects of the inefficiency of the original *h-index* are apparent; or to state it in its real dimension, significant efforts are needed to unfold the full potential of *h-index*. Firstly, the original *h-index* assigns the same importance to all citations, no matter what their age is, thus refraining from revealing the trendsetters scientists. Secondly, the *h-index* assigns the same importance to all articles, thus making the young researchers to have a relatively small *h-index*, because they did not have enough time either to publish a lot of good articles, or time to accumulate large number of citations. Thus, the *h-index* can not reveal the brilliant though young scientists.

After the introduction of the *h-index*, a number of other proposals followed, either presenting case studies using it (Bar-Ilan, 2006; Braun et al., 2005; Rousseau, 2006), or describing a new variation of it (Egghe, 2006*b*) (aiming to bridge the gap between the lower bound of total number of citations calculated by *h-index* and their real number), or studying its mathematics and its performance (Bornmann and Daniel, 2005; Egghe, 2006*a*). The interested reader can find a survey of the articles about *h-index* in Bornmann and Daniel (2007).

Deviating from their line of research, Sidiropoulos et al. (2007) developed a pair of generalizations of the *h-index* for ranking scientists, which are novel citation indices, a normalized variant of the *h-index* and a pair of variants of the *h-index* suitable for journal/conference ranking.

1.1.1 The contemporary h-index

The original *h-index* does not take into account the "age" of an article. It may be the case that some scientist contributed a number of significant articles that produced a large *h-index*, but now s/he is rather inactive or retired. Therefore, senior scientists, who keep contributing nowadays, or brilliant young scientists, who are expected to contribute a large number of significant works in the near future but now they have only a small number of important articles due to the time constraint, are not distinguished by the original *h-index*. Thus, arises the need to define a generalization of the *h-index*, in order to account for these facts.

We have defined a score $S_c(i)$ for an article i based on citation counting, as follows:

$$S_c(i) = \gamma * (Y(now) - Y(i) + 1)^{-\delta} * |C(i)| \tag{2}$$

where $Y(i)$ is the publication year of article i and $C(i)$ are the articles citing the article i. If we set δ=1, then $S_c(i)$ is the number of citations that the article i has received, divided by the "age" of the article. Since, we divide the number of citations with the time interval, the quantities $S_c(i)$ will be too small to create a meaningful *h-index*; thus, we use the coefficient γ. In the experiments reported by Sidiropoulos et al. (2007) the value of 4 is used for the coefficient γ and

the value of 1 for δ. In Section 3. we will use the same values. Thus, for an article published during the current year, its citations account four times. For an article published 4 year ago, its citations account only one time. For an article published 6 year ago, its citations account $\frac{4}{6}$ times, and so on.

This way, an old article gradually loses its "value", even if it still gets citations. In other words, in the calculations we mainly take into account the newer articles[2]. Therefore, we define a novel citation index for scientist rankings, the *contemporary h-index*, expressed as follows:

Definition 3 *A researcher has contemporary h-index h_c, if h_c of its N_p articles get a score of $S_c(i) \geq h_c$ each, and the rest $(N_p - h_c)$ articles get a score of $S_c(i) \leq h_c$.*

1.1.2 The trend h-index

The original *h-index* does not take into account the year when an article acquired a particular citation, i.e., the "age" of each citation. For instance, consider a researcher who contributed to the research community a number of really brilliant articles during the decade of 1960, which, say, got a lot of citations. This researcher will have a large *h-index* due to the works done in the past. If these articles are not cited anymore, it is an indication of an outdated topic or an outdated solution to the problem. On the other hand, if these articles continue to be cited, then we have the case of an *influential mind*, whose contributions continue to shape newer scientists' minds. There is also a second very important aspect in aging the citations. There is the potential of disclosing *trendsetters*, i.e., scientists whose work is considered pioneering and sets out a new line of research that currently is hot ("trendy"), thus this scientists' works are cited very frequently.

To handle this, we take the opposite approach than *contemporary h-index*'s; instead of assigning to each scientist's article a decaying weight depending on its age, we assign to each citation of an article an exponentially decaying weight, which is as a function of the "age" of the citation. This way, we aim at estimating the impact of a researcher's work in a particular time instance. We are not interested in how old the articles of a researcher are, but whether they still get citations. We define an equation similar to Equation 2, which is expressed as follows:

$$S_t(i) = \gamma * \sum_{\forall x \in C(i)} (Y(now) - Y(x) + 1)^{-\delta} \tag{3}$$

where γ, δ, $Y(i)$ and $S(i)$ for an article i are as defined earlier. We define a novel citation index for scientist ranking, the *trend h-index*, expressed as follows:

Definition 4 *A researcher has trend h-index h_t if h_t of its N_p articles get a score of $S_t(i) \geq h_t$ each, and the rest $(N_p - h_t)$ articles get a score of $S_t(i) \leq h_t$ each.*

Apparently, for $\gamma = 1$ and $\delta = 0$, the *trend h-index* coincides with the original *h-index*.

1.2 Our contributions

The purpose of our work is to extend and generalize the original *h-index* and its variations in such ways, so as to reveal various latent though strong facts hidden in citation networks. In this context, the article makes the following contributions:

[2] Apparently, if δ is close to zero, then the impact of the time penalty is reduced, and, for $\delta = 0$, this variant coincides with the original *h-index* for $\gamma = 1$.

- Introduces a generalization of the *h-index*, namely the *age decaying h-index*, which is appropriate for scientist ranking and is able to reveal *brilliant young scientists* and *trend-setters*. This generalization can also be used for conferences and journals ranking.
- Performs an extensive experimental evaluation of the aforementioned citation indices, using real data from DBLP, an online bibliographic database.

The rest of this article is organized as follows: In Section 2., we present the novel citation index *age decaying h-index*, and in Section 3. presents the evaluation of the introduced citation index against its predecessors. Finally, Section 4. summarizes the paper's contributions and concludes the article.

2. A NOVEL CITATION INDEX FOR SCIENTIST, CONFERENCES AND JOURNALS RANKING

2.1 The age decaying h-index

The *trend h-index* takes into account the "age" of the citations. On the on the hand *contemporary h-index* takes into account the "age" of the publications. The *age decaying h-index* is a generalization of both the *contemporary h-index* and *trend h-index*, which takes into account both the age of a scientist's article and the age of each citation to his/her articles.

We define a score function S_{ad} for a publication i as:

$$S_{ad}(i) = \gamma^2 * (Y(now) - Y(i) + 1)^{-\delta_1} * \sum_{\forall x \in C(i)} (Y(now) - Y(x) + 1)^{-\delta_2} \quad (4)$$

where γ, δ_1, δ_2 and $Y(i)$ for an article i are as defined earlier. If δ_1 and δ_2 are equal, then the "age" of the publication and the "age" of the citation have the same importance. We may give greater importance to one of them by increasing the corresponding δ (δ_1 or δ_2).

We define a novel citation index for scientist ranking, the *age decaying h-index*, expressed as follows:

Definition 5 *A researcher has age decaying h-index h_{ad} if h_{ad} of its N_p articles get a score of $S_{ad}(i) \geq h_{ad}$ each, and the rest $(N_p - h_{ad})$ articles get a score of $S_{ad}(i) \leq h_{ad}$ each.*

Likewise, the *age decaying h-index* can be defined for a Journal or a Conference. For instance, the *age decaying h-index* of a journal/magazine or a Conference is h_{ad}, if h_{ad} of the N_p articles that contains, have received at least h_{ad} citations each, and the rest $(N_p - h_{ad})$ articles received no more than h_{ad}.

The second metric of the original *h-index* notion is the factor a. Factor a_{ad} can be defined as:

$$\sum_{\forall i \in P} S_t(i) = a_{ad} * h_{ad}^2 \quad (5)$$

where P is the set of a scientist's publications. The a-index can be used as a second metric-index for the evaluation and ranking of scientists. It describes the age decaying "magnitude" of each scientist's "hits". A large a implies that some article(s) have received a fairly large number of citations compared to the rest of its articles and with respect to what the h-index presents.

3. EXPERIMENTS

Having defined this generalization and variants of the original *h-index*, we will evaluate in the subsequent sections their success in identifying scientists or forums with extraordinary performance or their ability to reveal latent facts in a citation network, such as brilliant young scientists and trendsetters. For the evaluation, we will exploit the on-line database of DBLP [3].

In the sequel, we will present a small subset of the results obtained for ranking scientists, conferences and journals, using the basic *h-index* definition as well as using the generalization developed in the previous section. Along the lines of (Sidiropoulos and Manolopoulos, 2005*a*,*b*, 2006), our dataset consists of the DBLP collection (DBLP timestamp: Mar/3/2006). The reason for selecting this source of data instead of ISI or Google data is twofold:

1. DBLP contains data about journal and conference publications as well, and
2. DBLP data are focused mostly in the area of Databases.

It is worthwhile noticing that many top conferences of this area are very competitive (with an acceptance ratio stronger than 1:3 and up to 1:7), and occasionally more competitive that the top journals of the area. In many computer science departments worldwide, publications in these conferences are favored in comparison to journal publications. Therefore, a ranking of conferences on databases is equally important to the ranking of the journals of the area.

The reason for selecting this "old" snapshot of the DBLP database is to be able to compare the results with our former published research. The used database snapshot contains 451694 inproceedings, 266307 articles, 456511 authors, 2024 conference series and 504 journals. Also, the number of citations in our dataset is 100205. Although this number is relatively small, it is a satisfactory sample for our purposes. Almost all citations in the database are made from publications prior to the year 2001. Thus, we can assume that the results presented here correspond to the year 2001. From now on, with the term "now" we actually mean sometime near 2001. Although other bibliographic sources, like ISI, are widely available and much more complete, the used collection has the two above desired characteristics and thus it is sufficient for exhibiting the benefits of our proposed citation indices, without biasing our results.

3.1 Experiments with the h-index for scientists

In Tables 1 and 2 we present the resulting ranking using the *h-index*, as well as its defined generalization, the *age decaying h-index*. In these tables column a_{ad} stands for the factor a of the *age decaying h-index*. Table 1 is sorted by the *h-index* ranking position. In this table we also present the values for *contemporary h-index* (h_c), *trend h-index* (h_t) and *age decaying h-index* (h_{ad}) and the corresponding rank position (sub-columns @ pos). For example, at the first position is ranked Michael Stonebraker with *h-index* 24, $a = 3.78$, total number of citations equal to 2180, total number of published papers =193, *age decaying h-index* equals 11 and his corresponding position at the *age decaying h-index* rank table is position number 14, *contemporary h-index* equals 13 and his position with the *contemporary h-index* metric is number 3, . . .

At a first glance, we see that the values computed for *h-index* (Table 1) are much lower than the values presented in (Hirsch, 2005) for physics scientists due to the non completeness of the

[3] The DBLP digital library with bibliographic data on "Databases and Logic Programming" is maintained by Michael Ley at the University of Trier, accessible from `http://dblp.uni-trier.de/`

Table 1. Scientist ranking with *h-index*.

Name	h	a	$N_{c,tot}$	N_p	h_{ad}(@ pos)	h_c(@ pos)	h_t(@ pos)
1.Michael Stonebraker	24	3.78	2180	193	11(@ 14)	13(@ 3)	19(@ 3)
2.Jeffrey D. Ullman	23	3.37	1783	227	14(@ 6)	14(@ 2)	20(@ 2)
3.David J. DeWitt	22	3.91	1896	150	14(@ 7)	16(@ 1)	23(@ 1)
4.Philip A. Bernstein	20	3.39	1359	124	7(@ 73)	10(@ 15)	12(@ 23)
5.Won Kim	19	2.96	1071	143	7(@ 71)	10(@ 12)	14(@ 12)
6.Catriel Beeri	18	3.16	1024	93	7(@ 66)	10(@ 13)	13(@ 18)
7.Rakesh Agrawal	18	3.06	994	154	16(@ 1)	13(@ 4)	19(@ 4)
8.Umeshwar Dayal	18	2.81	913	130	8(@ 45)	9(@ 20)	13(@ 16)
9.Hector Garcia-Molina	17	3.60	1041	314	13(@ 9)	10(@ 8)	17(@ 7)
10.Yehoshua Sagiv	17	3.52	1020	121	9(@ 35)	9(@ 18)	13(@ 14)
11.Ronald Fagin	17	2.83	818	121	5(@ 130)	7(@ 48)	11(@ 38)
12.Jim Gray	16	6.13	1571	118	11(@ 16)	11(@ 7)	14(@ 10)
13.Serge Abiteboul	16	4.33	1111	172	16(@ 3)	12(@ 5)	17(@ 6)
14.Michael J. Carey	16	4.25	1090	151	10(@ 22)	10(@ 9)	14(@ 11)
15.Nathan Goodman	16	3.37	865	68	5(@ 161)	7(@ 49)	10(@ 49)
16.Christos Faloutsos	16	2.89	742	175	13(@ 10)	10(@ 11)	17(@ 8)
17.Raymond A. Lorie	15	6.23	1403	35	5(@ 134)	8(@ 29)	11(@ 33)
18.Jeffrey F. Naughton	15	2.90	653	123	14(@ 8)	10(@ 10)	15(@ 9)
19.Bruce G. Lindsay	15	2.76	623	60	6(@ 91)	8(@ 37)	12(@ 32)
20.David Maier	14	5.56	1090	158	8(@ 49)	10(@ 14)	12(@ 24)

Table 2. Scientist ranking with *age decaying h-index*.

Name	h_{ad}	a_{ad}	$N_{c,tot}$	N_p	h(@ pos)	h_c(@ pos)	h_t(@ pos)
1.Rakesh Agrawal	16	3.28	994	154	18(@ 7)	13(@ 4)	19(@ 4)
2.Jennifer Widom	16	3.19	709	136	14(@ 23)	12(@ 6)	18(@ 5)
3.Serge Abiteboul	16	3.08	1111	172	16(@ 13)	12(@ 5)	17(@ 6)
4.Dan Suciu	16	2.79	244	113	9(@ 100)	9(@ 22)	12(@ 25)
5.Alon Y. Levy	15	2.85	321	77	10(@ 69)	9(@ 21)	14(@ 13)
6.Jeffrey D. Ullman	14	4.18	1783	227	23(@ 2)	14(@ 2)	20(@ 2)
7.David J. DeWitt	14	3.41	1896	150	22(@ 3)	16(@ 1)	23(@ 1)
8.Jeffrey F. Naughton	14	2.95	653	123	15(@ 18)	10(@ 10)	15(@ 9)
9.Hector Garcia-Molina	13	4.07	1041	314	17(@ 9)	10(@ 8)	17(@ 7)
10.Christos Faloutsos	13	2.62	742	175	16(@ 16)	10(@ 11)	17(@ 8)
11.Daniela Florescu	13	2.44	105	60	5(@ 324)	8(@ 43)	9(@ 69)
12.Hans-Peter Kriegel	12	3.26	465	204	11(@ 50)	8(@ 28)	12(@ 21)
13.Joseph M. Hellerstein	12	2.76	252	86	10(@ 79)	8(@ 36)	12(@ 31)
14.Michael Stonebraker	11	4.12	2180	193	24(@ 1)	13(@ 3)	19(@ 3)
15.H. V. Jagadish	11	3.59	503	151	12(@ 39)	10(@ 16)	13(@ 17)
16.Jim Gray	11	3.58	1571	118	16(@ 12)	11(@ 7)	14(@ 10)
17.Surajit Chaudhuri	11	3.22	263	114	9(@ 97)	8(@ 34)	12(@ 30)
18.Yannis Papakonstantinou	11	3.06	219	57	8(@ 124)	8(@ 39)	10(@ 48)
19.Tova Milo	11	2.53	179	74	8(@ 133)	8(@ 41)	9(@ 64)
20.Leonid Libkin	11	2.46	143	99	6(@ 248)	6(@ 78)	10(@ 52)

source data. Also, we can notice that the values for h and h_{ad} are different with each other as well as there are differences in the ordering of the scientists. This confirms our allegation for the convenience of these indices.

In contrast with our contemporary and trend h-index research (Sidiropoulos et al., 2007), Tables 1 and 2 present significant differences. The rank order of Table 1 is expected, since well known names of the database domain are ranked at the first 20 positions. On the other hand, Table 2 presents a different ordering with new names appeared at the first 20 positions. The researchers that are reported to be in the top 20 with the *age decaying h-index* but not with the original *h-index* are: Dan Suciu, Alon Y. Levy, Daniela Florescu, Hans-Peter Kriegel, Joseph Hellerstein, H. V. Jagadish, Surajit Chaudhuri, Yannis Papakonstantinou, Tova Milo and Leonid Libkin. Also, the ordering given by *age decaying h-index* is different than the ones of *contemporary h-index* and *trend h-index*. This fact confirms that the *age decaying h-index* is a

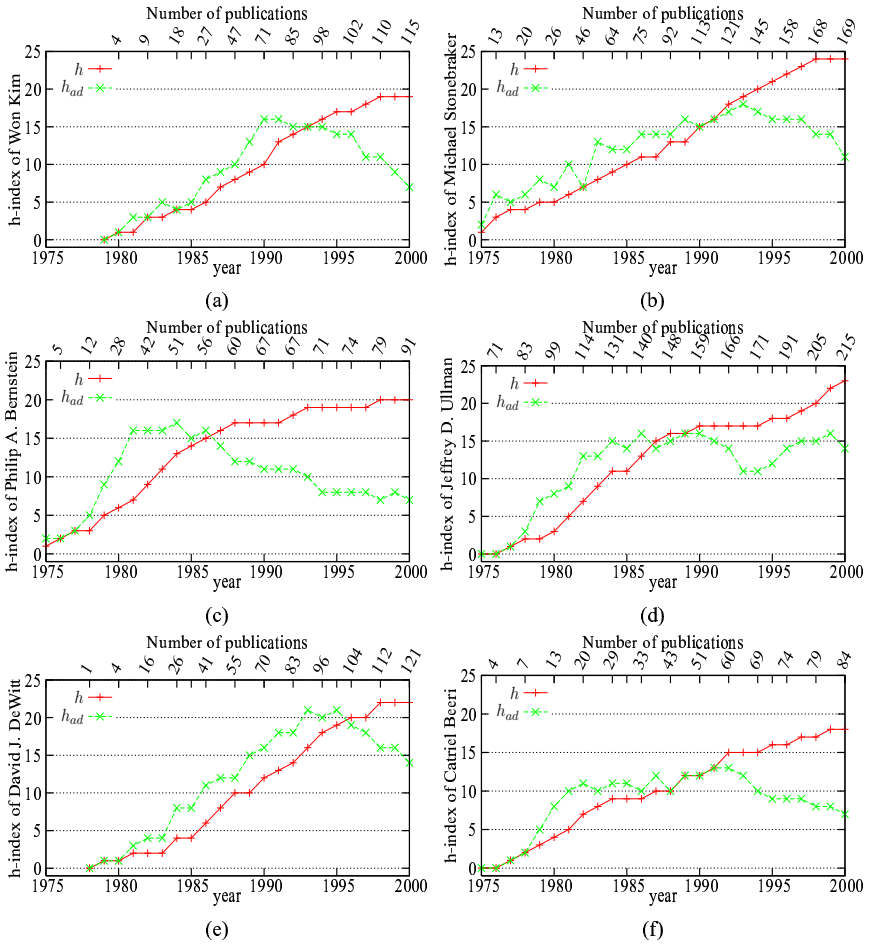

Figure 1. The *h-index* of scientists working in databases area.

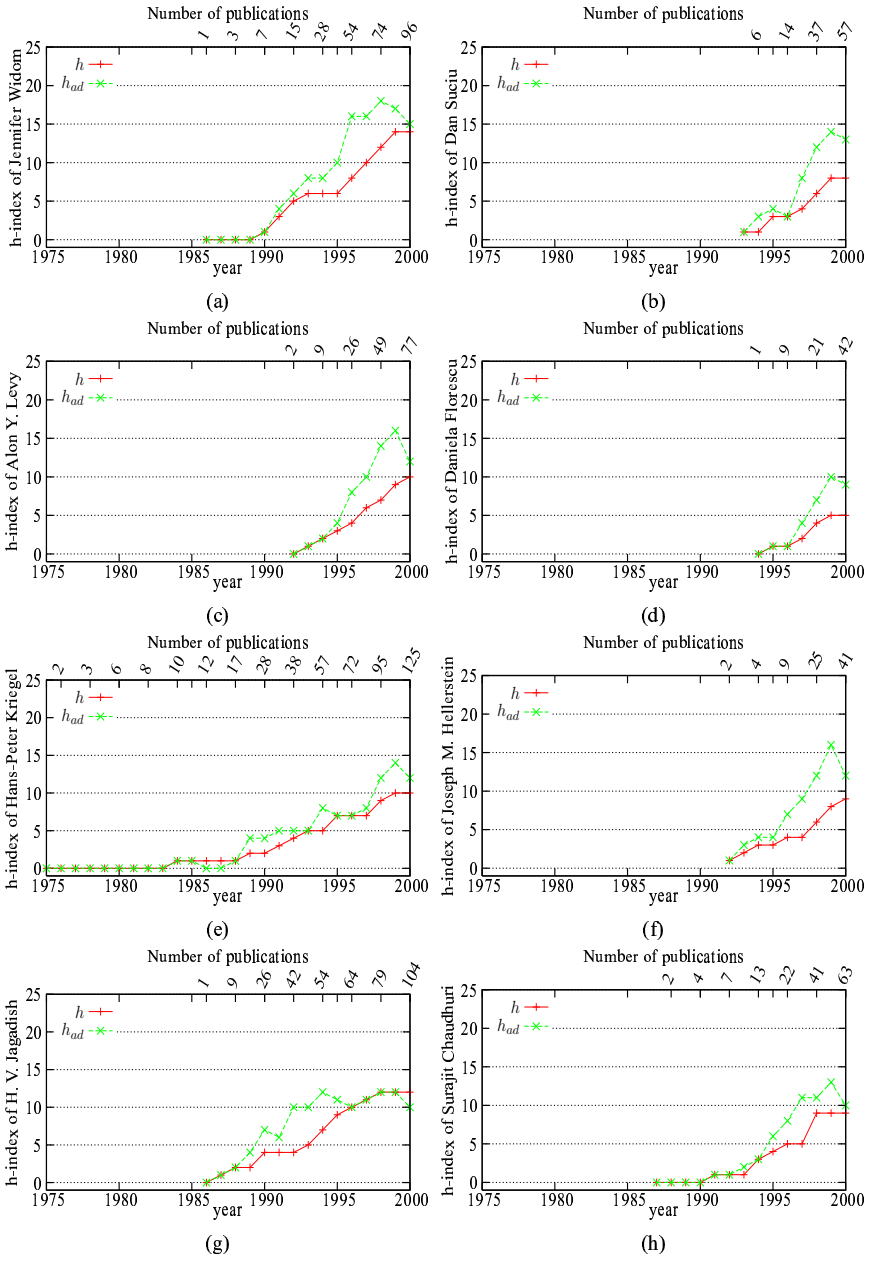

Figure 2. The *h-index* of scientists working in databases area (part b).

novel method. The majority of the "new" scientists at the top 20 positions, can be said that are "young" scientists compared to the "all time classics" scientists. This can also be confirmed from the Figures 2 and 1. As we can see in Figure 2 most of the "new" ones have started published

around 1990, in contrast with the scientists presented in Figure 1 who they started publishing around 1975. This means that our new index is really age decaying. Thus, it assists the scientists with new publications and simultaneously new citations.

It is also worthwhile to mention that the *contemporary h-index* and *trend h-index* are fair metrics for the "all-time classic" scientists, e.g., Jeffrey Ullman, Michael Stonebraker, and David DeWitt, whose influential works continue to shape the modern scientists way of thinking.

Motivated by the differences in the above tables, we present the collection of graphs in Figure 1. In these figures, we can see the history of the *h-index* for those scientists, who present significant differences between the *h-index* family of citation indices, and also those who have a rapid upward slope at their plot curves. Again, we remind that our data set is rather incomplete for the years after 2000, and thus a downwards pitch for all the researchers appears during the years 1999-2000. However, the results are indicative.

Won Kim (Figure 1(a)), Michael Stonebraker (Figure 1(b)) and Philip A. Bernstein (Figure 1(c)) present a similar path. For instance, there is a high ascending curve for *age decaying h-index* until around 1990 (with few years difference). Therefore, we expect that *h-index* will not present high increase. This is explained by the fact that the main research interests of Won Kim was on object-oriented database systems, which flourished during the last years of the eighties and in the first years of the nineties, but later become a relatively inactive area. Stonebraker and Bernstein, after their intensive and high quality research, which laid the foundations of the relational model during the '80s, reduced their productivity.

Jeffrey D. Ullman's h_{ad} followed an uprising course until 1985, then started to be stabilized and lightly decreasing, but after 1994 it started increasing again. This is due to the fact that at that time, J. D. Ullman worked with his colleagues on the integration of distributed data sources and particularly his research focused on semistructured data, that happened to be very popular and trendy research theme.

The pattern of increase of the age decaying h-index for David DeWitt and Catriel Beeri is quite similar, with a shift of a few years in the time scale, both of whom, after fundamental contributions to the theory and practice of the relational model that brought them at the forefront of the research, did not dealt with the new research topics that emerged at that time.

In Figure 2(a), we see the progress rate for Jennifer Widom. While Jennifer Widom is not even among the top 20 researchers using the *h-index*, she is on the 2^{nd} position using the *age decaying h-index*. Also, she is ranked 6^{th} and 5^{th} using the *contemporary h-index* and *trend h-index*, respectively. She is one of the researchers from our list that presents such a big difference on the timing rate compared to the basic *h-index*. As we can also see from the diagram, this difference is justifiable, since the increase rate of the basic *h-index* is high. Jenifer Widom made some ground breaking contributions on building semistructured data management systems, that laid the foundations for the modern XML management systems.

Dan Suciu climbed from the 100^{th} place by the original *h-index* to the 4^{th} by the *age decaying h-index*. Figure 2(b) shows that the *age decaying h-index* follows a rapidly ascending course, as well as that for Alon Y. Levy presented in Figure 2(c). Daniela Florescu gained the highest rise from all the scientists presented in this paper. She is ranked at the 324^{th} place by the original *h-index* and moved to the 11^{th} position. The pattern of growth of all these scientists is not accidental; all of them have worked on the topic of semistructured data, which later was transformed to the area of XML data management, which can be easily recognized as one of the most hot and trendy topics during the last years of the previous decade and the first years of this

decade.

Joseph M. Hellerstein (Figure 2(f)) and Surajit Chaudhuri (Figure 2(h)) follow a similar slope. Although both researchers have broad research interests, it is easy to ascribe the growth of their *age decaying h-index* to their contributions to the relational databases and to online analytical processing (OLAP) and data warehousing.

Hans-Peter Kriegel (see Figure 2(e)) has been recognized as one of the most productive researchers in the area of spatial data management; this topic was very popular and attracted a lot of interest during the previous decade. Therefore, the pattern of growth of his *age decaying h-index* is reasonable. Similarly, the *age decaying h-index* of H. V. Jagadish, who was working at that time on multidimensional data, exhibits similar growth pattern.

Collectively, starting from the observations about the scientists with steep growing of their *age decaying h-index*, we can go one step further and recognize research topics which constitute the preferred and trendy research areas at that periods, like spatial data, semistructured data and OLAP. Indeed, the findings of our citations indexes are in absolute accordance with what the common sense deduces by observing the number of paper on each topic in major journals and conferences. Thus, the proposed citation index is able to reveal large scientific areas as promising topics for research.

3.2 Experiments with conferences and journals ranking

3.2.1 Experiments with conferences ranking

To evaluate our citation indices on conference ranking, we extract only the database conferences (as defined by Elmacioglu and Lee (2005)) from the data we used in the previous section. In this section we will make experiments using the indicator that we fixed for scientists, namely *h-index* and *age decaying h-index*.

In Table 3 we present the top-10 conferences using the *h-index* for the ordering. The symbol a in Table 3 and the symbol a_{ad} in Table 4 correspond to the *a-index* on Definition 2 and Equation 5 respectively. Since the quality of the conferences is relatively constant, we observe that in Tables 3 and 4 there are no significant differences in the ranking. The differences occur below the 5^{th} place where "International Conference on Conceptual Modeling (ER)" and "Expert Database Systems (EDS)" are replaced by "International Conference on Database Theory (ICDT)" and "Knowledge Discovery and Data Mining (KDD)".

In Figure 3 we present in the same way we used for scientists, the progress of selected conferences. Note here that the *h-index* is shown per year in the graphs, which means that this is the computed *h-index* during the specific year. E.g., the *h-index* that is computed for the VLDB for 1995 is the *h-index* that is computed if we exclude everything from our database after 1995.

Due to the lack of citations for the years after 1999, in all graphs there is a stabilization of the *h-index* line and a downfall for the indicator *age decaying h-index*. Figure 3(a) presents the history of the SIGMOD conference. According to the tables, SIGMOD is ranked first. In the figure, we observe its steeply ascending line as well as the *age decaying h-index* remains higher than the *h-index* (until 1999). Also, VLDB (Figure 3(b)) follows an ascending path. These two conferences are clearly ranked first by our algorithm and by *h-index*. On the other hand, the PODS conference (Figure 3(c)) follows a bending line after 1988 with some picks. ICDE is a relatively younger conference compared to the rest of the conferences presented, but we can see in the plot (Figure 3(d)), that it follows a rapidly ascending course until 1987 and afterwords it's

age decaying h-index is almost stabilized with an increasing trend.

Finally, with respect to the ADBT conference (Figure 3(e)) we mention that this conference was organized only three times (1977, 1979 and 1982). As we can see in the upper x axis, the number of publications stops increasing after 1982. Thus, we can not compare it to the rest of the conferences. What we observe from this plot, is that the *age decaying h-index* converges to zero which confirms the correctness of our metric.

KDD is the "youngest" conference among the rest, but it has managed to climb up to the 6^{th} place in the *age decaying h-index* rank table. From the plot (Figure 3(f)) we cannot gather much more information due to its short history and the lack of available data.

3.2.2 Experiments with journals ranking

In the case of journals, we can use the basic form of *h-index* as well as the generalization *age decaying h-index* we defined for scientists and for conferences.

Tables 5 and 6 present the top-10 journals according to the aforementioned indices. As expected, the ACM TODS (tods), IEEE TKDE (tkde), SIGMOD Record (sigmod) are the top three journals. The striking observation is that the Information Systems (is) drops in the ranking with the *age decaying h-index*, as compared to its position with *h-index*, implying that it is not considered a prestigious journal anymore; it is ranked even below the Data Engineering Bulletin!,

Table 3. Conferences ranking with *h-index*.

Name	h	a	$N_{c,tot}$	N_p
1.sigmod	45	6.05	12261	2059
2.vldb	37	7.10	9729	2192
3.pods	26	5.74	3883	776
4.icde	22	6.83	3307	1970
5.er	16	5.80	1486	1338
6.edbt	13	3.89	658	434
7.eds	12	3.65	527	101
8.adbt	12	2.86	412	42
9.icdt	11	4.79	580	313
10.oodbs	11	3.96	480	122

Table 4. Conferences ranking with *age decaying h-index*.

Name	h_{ad}	a_{ad}	$N_{c,tot}$	N_p	h
1.sigmod	32	5.85	12261	2059	45
2.vldb	25	6.94	9729	2192	37
3.pods	20	5.32	3883	776	26
4.icde	17	8.01	3307	1970	22
5.icdt	12	5.27	580	313	11
6.kdd	11	4.08	243	1074	6
7.edbt	11	3.92	658	434	13
8.webdb	9	2.69	31	163	3
9.cikm	8	4.18	211	1030	7
10.ssdbm	8	3.71	321	609	7

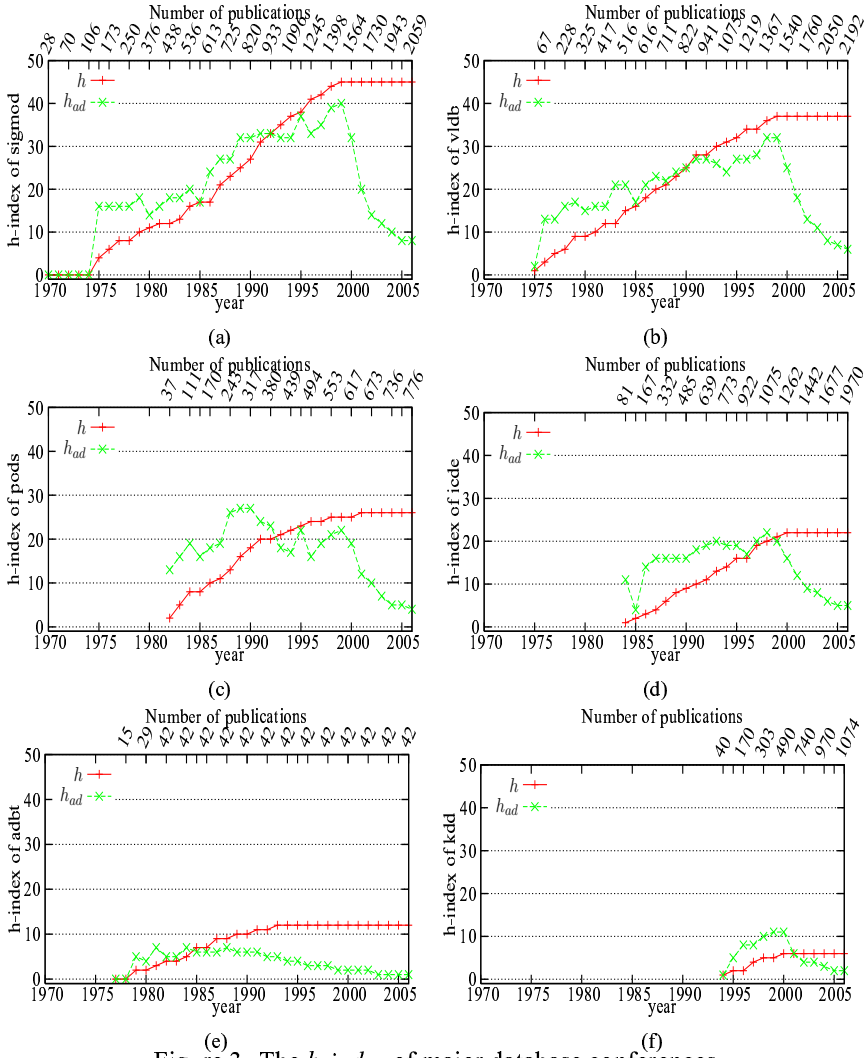

Figure 3. The *h-index* of major database conferences.

implying that it is not publishing modern research results as it used to. On the contrary, SIGMOD Record and VLDB Journal (vldb) show an uprising trend.

In Figure 4 we present the results of computing the defined indices for the major journals of the database domain on a per year basis. Due to the lack of available data after the year 2000, all indices drop steeply. Though, the case of ACM TODS is worthwhile mentioning. Its *age decaying h-index* (Figure 4(a)) drops after 1993, which can be attributed to the relatively large end-to-end publication time of its articles during the years 1990-2000 (Snodgrass, 2003), which acted as an impediment for the authors to submit their works in that venue. Fortunately, this is not the case anymore. On the other hand, SIGMOD Record (Figure 4(c)) and VLDB Journal (Figure 4(d)) show a clear uprising trend until 1998. Also, the case of SIGMOD Record is

characteristic, because, even though it has been published since 1970, its indices get really noticeable only after 1980, when this newsletter started to publish some very good survey-type articles and was freely available on the Web, which improved its visibility. Finally, Information Systems (is: Figure 4(e)) and ACM Transactions on Information Systems (tois: Figure 4(f)) show a stable performance based on the *age decaying h-index* (of course by ignoring the years after 1999 due to the lack of data).

4. CONCLUSIONS

Estimating the significance of a scientist's work is a very important issue for prize awarding, faculty recruiting; similarly, the estimation of a publication forum's (journal or conference) is significant since it impacts the scientists' decisions about where to publish their work. This issue has received some attention during the last years, but the interest on this topics has been renewed by a path-breaking paper by J. E. Hirsch, who proposed the *h-index* to perform fair ranking of scientists, avoiding many of the drawbacks of the earlier bibliographic ranking methods.

The initial proposal and meaning of the *h-index* has various shortcomings, mainly of its inability to differentiate between active and inactive (or retired) scientists and its weakness to

Table 5. Journal ranking with *h-index*.

Name	h	a	$N_{c,tot}$	N_p
1.tods	49	3.88	9329	598
2.tkde	18	4.69	1520	1388
3.is	16	4.71	1208	934
4.sigmod	15	5.07	1142	1349
5.tois	13	4.37	740	378
6.debu	11	7.13	863	877
7.vldb	9	5.03	408	281
8.ipl	8	6.06	388	4939
9.dke	6	8.77	316	773
10.dpd	6	5.25	189	238

Table 6. Journal ranking with *age decaying h-index*.

Name	h_{ad}	a_{ad}	$N_{c,tot}$	N_p	h
1.tods	13	7.71	9329	598	49
2.sigmod	13	4.94	1142	1349	15
3.tkde	12	5.77	1520	1388	18
4.debu	12	3.49	863	877	11
5.vldb	12	2.82	408	281	9
6.dpd	7	3.82	189	238	6
7.is	6	7.51	1208	934	16
8.jiis	6	5.67	156	318	6
9.tois	5	7.14	740	378	13
10.dke	5	6.52	316	773	6

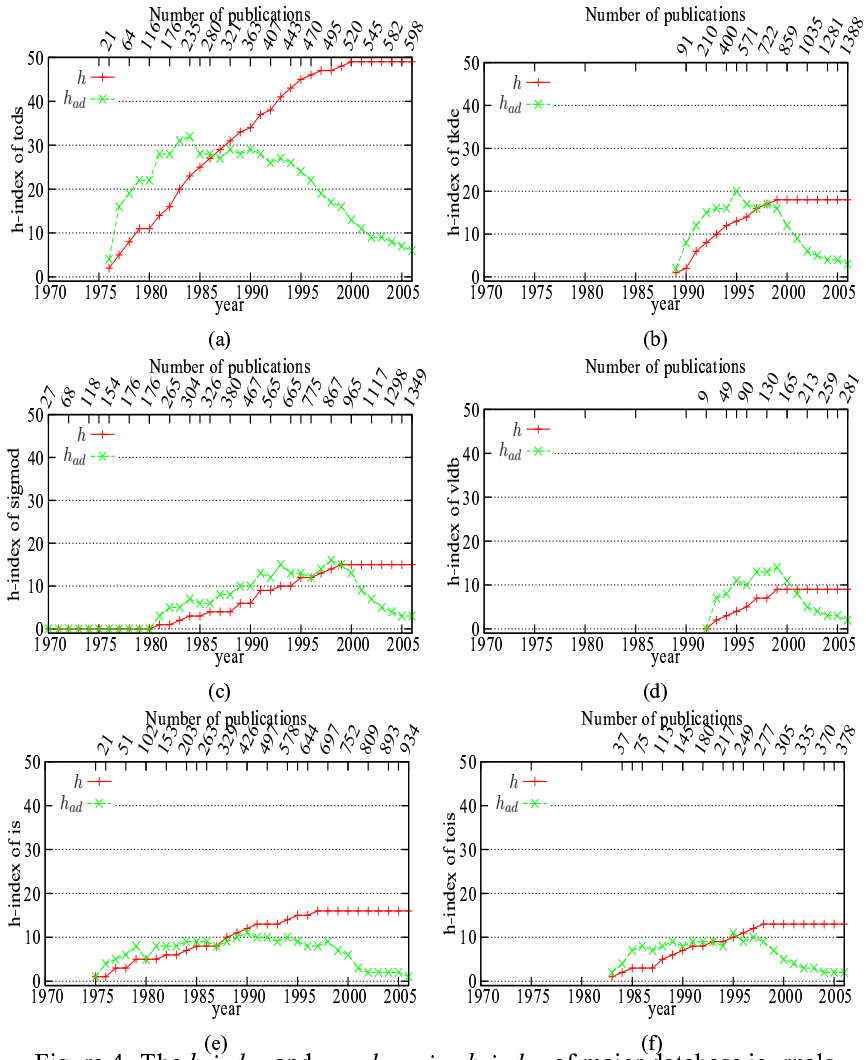

Figure 4. The *h-index* and *age decaying h-index* of major database journals.

differentiate between significant works in the past (but not any more) and the works which continue to shape the scientific thinking.

Based on the identification of these shortcomings of *h-index*, we proposed in this article an effective *age decaying h-index* generalization. This novel citation index aim at the ranking of scientists by taking into account both the age of the published articles as well as the age of the citations to each article.

To evaluate the proposed ranking metrics, we conducted extensive experiments on an online bibliographic database containing data from journal and conference publications as well, and moreover focused in the specific area of databases. From the results we obtained, we concluded that *h-index* is not a general purpose indicative metric. The *age decaying h-index* is able to

disclose latent facts in citation networks, like trendsetters and brilliant young scientists. For the case of conference and journal ranking, the index *age decaying h-index* gives a more fair view for the ranking.

REFERENCES

Ball, P. (2005), "Index aims for fair ranking of scientists – h-index sums up publication record", *Nature* , Vol. 436, p. 900.

Bar-Ilan, J. (2006), "h-index for price medalists revisited", *ISSI Newsletter* , Vol. 5.

Barnes, S. J. (2005), "Assessing the value of IS journals", *Communications of the ACM* , Vol. 48, pp. 110–112.

Bernstein, P. A., Bertino, E., Heuer, A., Jensen, C. J., Meyer, H., Tamer Ozsu, M., Snodgrass, R. T. and Whang, K.-Y. (2005), "An apples-to-apples comparison of two database journals", *ACM SIGMOD Record* , Vol. 34, pp. 61–64.

Bharati, P. and Tarasewich, P. (2002), "Global perceptions of journals publishing e-commerce research", *Communications of the ACM* , Vol. 45, pp. 21–26.

Bornmann, L. and Daniel, H.-D. (2005), "Does the h-index for ranking of scientists really work?", *Scientometrics* , Vol. 65, pp. 391–392.

Bornmann, L. and Daniel, H.-P. (2007), "What do we know about the h-index?", *Journal of the American Society of Information Science and Technology* . to appear.

Braun, T., Glanzel, W. and Schubert, A. (2005), "A Hirsch-type index for journals", *The Scientist* , Vol. 19, pp. 8–10.

Egghe, L. (2006*a*), "Dynamic h-index: The Hirsch index in function of time", *Scientometrics* . to appear.

Egghe, L. (2006*b*), "Theory and practise of the g-index", *Scientometrics* , Vol. 69, pp. 131–152.

Elmacioglu, E. and Lee, D. (2005), "On six degrees of separation in DBLP-DB and more", *ACM SIGMOD Record* , Vol. 34, pp. 33–40.

Garfield, E. (1972), "Citation analysis as a tool in journal evaluation", *Science* , Vol. 178, pp. 471–479.

Hirsch, J. E. (2005), "An index to quantify an individual's scientific research output", *Proceedings of the National Academy of Sciences* , Vol. 102, pp. 16569–16572.

Katerattanakul, P., Han, B. T. and Hong, S. (2003), "Objective quality ranking of computing journals", *Communications of the ACM* , Vol. 46, pp. 111–114.

Kelly Rainer, R. and Miller, M. D. (2005), "Examining differences across journal rankings", *Communications of the ACM* , Vol. 48, pp. 91–94.

Lowry, P., Romans, D. and Curtis, A. (2004), "Global journal prestige and supporting disciplines: A scientometric study of information systems journals", *Journal of the Association for Information Systems* , Vol. 5, pp. 29–75.

Mylonopoulos, N. A. and Theoharakis, V. (2001), "Global perception of IS journals", *Communications of the ACM* , Vol. 44, pp. 29–33.

Nascimento, M., Sander, J. and Pound, J. (2003), "Analysis of SIGMOD's co-authorship graph", *ACM SIGMOD Record* , Vol. 32, pp. 8–10.

Nerur, S. P., Sikora, R., Mangalaraj, G. and Balijepally, V. (2005), "Assessing the relative influence of journals in a citation network", *Communications of the ACM* , Vol. 48, pp. 71–74.

Rahm, E. and Thor, A. (2005), "Citation analysis of database publications", *ACM SIGMOD Record* , Vol. 34, pp. 48–53.

Rousseau, R. (2006), "A case study: Evolution of JASIS' Hirsch index", *Library and Information Science* . http://eprints.rcils.org/archive/00005430.

Schwartz, R. B. and Russo, M. C. (2004), "How to quickly find articles in the top IS journals", *Communications of the ACM* , Vol. 47, pp. 98–101.

Sidiropoulos, A., Katsaros, D. and Manolopoulos, Y. (2007), "Generalized Hirsch h-index for disclosing latent facts in citation networks", *Scientometrics* , Vol. 72. to appear.

Sidiropoulos, A. and Manolopoulos, Y. (2005*a*), "A citation-based system to assist prize awarding", *ACM SIGMOD Record* , Vol. 34, pp. 54–60.

Sidiropoulos, A. and Manolopoulos, Y. (2005*b*), "A new perspective to automatically rank scientific conferences using digital libraries", *Information Processing & Management* , Vol. 41, pp. 289–312.

Sidiropoulos, A. and Manolopoulos, Y. (2006), "Generalized comparison of graph-based ranking algorithms for publications and authors", *Journal for Systems and Software* , Vol. 79, pp. 1679–1700.

Snodgrass, R. (2003), "Journal relevance", *ACM SIGMOD Record* , Vol. 32, pp. 11–15.