Scientometrics (2011) 86:671–686
DOI 10.1007/s11192-010-0302-9

f-Value: measuring an article's scientific impact

Eleni Fragkiadaki · Georgios Evangelidis ·
Nikolaos Samaras · Dimitris A. Dervos

Received: 5 June 2010 / Published online: 26 October 2010

Abstract The f-value is a new indicator that measures the importance of a research article by taking into account all citations received, directly and indirectly, up to depth *n*. The f-value considers all information present in a Citation Graph in order to produce a ranking of the articles. Apart from the mathematical equation that calculates the f-value, we also present the corresponding algorithm with its implementation, plus an experimental comparison of f-value with two known indicators of an article's scientific importance, namely, the number of citations and the Page Rank for citation analysis. Finally, we discuss the similarities and differences among the indicators.

Keywords Citation analysis · Citation graph · f-Value · Page Rank

Introduction

The use of citation analysis has grown in importance during the past few years. The vast increase of scientific production made it very difficult for scientists to keep track of publications they might be interested in. Many indicators have been developed to rank scientific journals, authors and scientific publications by measuring their importance.

E. Fragkiadaki (✉) · G. Evangelidis · N. Samaras
Department of Applied Informatics, University of Macedonia Economic and Social Sciences,
54006 Thessaloniki, Greece
e-mail: eleni.fra@gmail.com

G. Evangelidis
e-mail: gevan@uom.gr

N. Samaras
e-mail: samaras@uom.gr

D. A. Dervos
Department of Information Technology, Alexander Technology Educational Institute (ATEI)
of Thessaloniki, 57400 Sindos, Greece
e-mail: dad@it.teithe.gr

The most widely used ranking indicator for journals is the Impact Factor proposed by Garfield (1955, 1999, 2005). The ranking is based on the average number of citations received per citable item in the journal in question during a predefined period of time (the past 2 years).

In order to measure the importance of a researcher's work, other metrics have been proposed that use the collection of all articles a researcher has (co-) authored, plus the sum of all direct citations received. Such indexes are the h-Index (Hirsch 2005), g-Index (Egghe 2006), and their variations.

For example, there have been variations of the h-index that take into account: (a) the total number of citations included in the Hirsch-core (A-index, R-index) (Jin et al. 2007), (b) the age of the publications included in the Hirsch-core (AR-index) (Jin et al. 2007), (c) the age of the publications of an author (contemporary h-index) (Sidiropoulos et al. 2007), (d) the age of the citations (trend h-index) (Sidiropoulos et al. 2007), (e) the combination of the above two (age-decaying h-index) (Katsaros et al. 2007), and, (f) not only the citations inside the Hirsch-core but also the ones received by publications currently not included in the Hirsch-core (tapered h-index) (Anderson et al. 2008).

There have been some variations of the g-index as well, like the gr-index and the grat-index (Guns and Rousseau 2009).

The importance of a scientific publication is most commonly measured based on the number of citations it has received. A different approach was proposed by Rousseau (Rousseau 1987), who claims that publications mentioned in the reference list have an impact on the publication in question, and also, recently, there has been a proposal for applying the philosophy of Page Rank (Brin and Page 1998) on a Citation Graph (Ma et al. 2008). Finally, the Cascading Citations Indexing Framework approach (Dervos and Kalkanis 2005; Dervos et al. 2006; Dervos and Klimis 2008) suggests that citations should be addressed at the (article, author) level in order to rank the contribution of each author's scientific work.

We suggest a new indicator for measuring the importance of a research article, the f-value. We produce a ranking of the publications included in the CiteSeer bibliographic database (Citeseer 1997; Giles et al. 1998) and compare our results with the ones obtained by other indicators.

In "Related work" section the Number of Citations, the Cascading Citations Indexing Framework, and the Page Rank for citation graphs approaches are presented. "f-Value description" section describes the basic concept of the f-value and in "Determining the reducing factor" section, we justify the selection of the specific reducing factor used in the calculation of the f-value. The paper continuous by presenting the f-value algorithm in " f-Value algorithm" section and the different rankings produced by three different indicators in "Experimental results" section. "Discussion" section describes the similarities and differences of the f-value with the other indicators, and, finally, the last section concludes the paper.

Related work

A citation graph is a representation of the relationships that exist between research articles based on the references that each article provides. In Fig. 1, articles are shown as nodes of a directed graph. In this example there are seven articles labeled A to G.

The arcs of the graph represent references among articles. For example, the arc leaving node B can be interpreted as "article B references article D". The incoming arcs are the

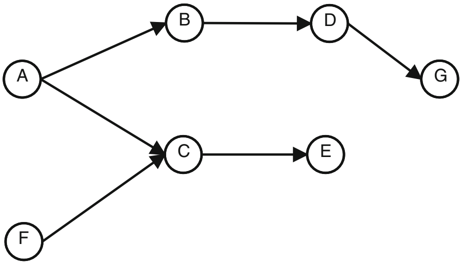


Fig. 1 Citation Graph 1

direct citations received by a specific article. For article D we can state that "article D receives one direct citation from article B".

Number of citations

This approach produces a ranking of scientific publications based on the number of citations they receive. It is by far the most simplistic approach, but, it is widely used. For example, in the citation graph of Fig. 1, articles A and F receive zero citations, articles B, D, E and G receive one citation each, and article C receives two citations.

The cascading citations indexing framework (c^2-IF)

The fundamental concept in the c^2-IF approach (Dervos and Kalkanis 2005; Dervos et al. 2006) is the *n*-gen citation. According to c^2-IF, direct citations like the ones discussed in the previous section are called 1-gen citations. If we carefully examine the citation graph in Fig. 1, we observe that article D also receives an indirect citation from article A, via article B. This is considered to be a 2-gen citation. In general, an *n*-gen citation exists between a source article S and a target article T, if there is a directed path in the citation graph from node S to node T. In the example of Fig. 1, the highest *n*-gen citation present is of depth 3: the one from article A to article G, along the citation path $A \rightarrow B \rightarrow D \rightarrow G$.

According to c^2-IF, the citations that a (article, author) pair receives can be calculated up to depth *n*, thus, producing a number of distinct values. So, if we choose to consider the citations up-to depth 3, the following values will be calculated: 1-gen citations, 2-gen citations, and 3-gen citations. These values are stored in a table called Medal Standings Output (MSO).

We also stress that the c^2-IF approach is not to be considered as a ranking method but merely a framework that extends the citation indexing paradigm to include 2-,3-,..., *k*-gen citations. We should also point out that in the c^2-IF approach, k is predefined and its value can range from $2 \ldots n$, where *n* is the maximum path present in the specific citation graph. In other words, $k \in [2 \ldots n]$ and consequently that many distinct values are going to be calculated for each article in the citation graph.

Page rank

The original Page Rank (Brin and Page 1998) produces a ranking of web pages by taking into account the number and importance of pages linking to each web page. The formula used by the Page Rank algorithm is

$$PR(A) = (1 - d) + d * \sum_i \frac{PR(T_i)}{C(T_i)} \tag{1}$$

where $PR(T_i)$ is the Page Rank value of page T_i linking to page A whose Page Rank value we wish to calculate, and $C(T_i)$ is the number of outbound links of page T_i. Finally, d is the damping factor. In order to better explain the damping factor, we should first give a general description of the concept of Page Rank.

The Page Rank algorithm is based on the Random Surfer model which states that a person, the "random surfer", navigates through the web randomly, by clicking on links present on a web page. So, how high a web page ranks has to do with the probability that this "random surfer" eventually visits the web page in question. The probability increases as the number of incoming links increases and the effect is even more intense if these links come from web pages which score high, thus having themselves high probability to be visited. But, there is always a chance that our "random surfer" gets bored and chooses to simply leave, a reaction indicated by the damping factor, which on the original article was chosen to be 0.85. In most discussions about Page Rank, 0.85 is the value used for the damping factor, but, there is at least one article that we know of that examines the behavior of the original Page Rank algorithm when different values are chosen (Boldi et al. 2009). So, for the most common value of the damping factor, Eq. 1 actually becomes

$$PR(A) = 0.15 + 0.85 * \sum_i \frac{PR(T_i)}{C(T_i)} \tag{2}$$

In (Ma et al. 2008) a variation of the original Page Rank algorithm is applied to citation graphs. In that article, the authors apply Eq. 1 by choosing d = 0.5. They choose the specific value based on an empirical study that states that researchers will probably not follow six articles and stop but only two.

f-Value description

The Cascading Citations Indexing Framework introduces the *k*-gen (indirect) citations as a means of acknowledging the importance of a research article based not only on its direct influence (number of 1-gen citations) but also on the influence the citing articles represent in their scientific field.

In this paper, we introduce the f-value, a new indicator that quantifies the importance of a research article. The f-value considers the accumulated importance of all articles that have based their scientific contribution on the article in question, directly or indirectly. In other words, each article's importance is represented by a single value, the f-value. The method used to calculate the f-values of articles in a citation graph is based on our complete knowledge of the graph, thus it is exhaustive in nature and considers all citation paths present up to the maximum depth *n*.

Let us consider the following example. We have six articles, labeled A to F related as shown in Fig. 2, thus producing the MSO table shown in Table 1.

A possible way to calculate the f-value of an article A by taking into account the indirect citations could be

$$f(A) = 1 + (f(A_1) + f(A_2) + \cdots + f(A_m)) \tag{3}$$

Fig. 2 Citation Graph 2

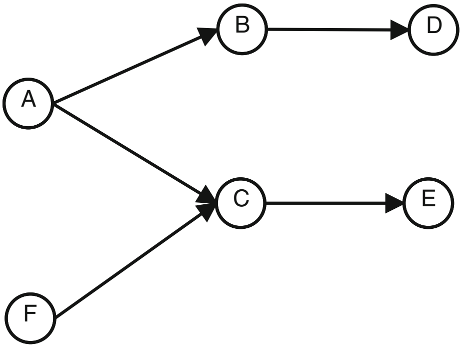


Table 1 MSO table for Citation Graph 2

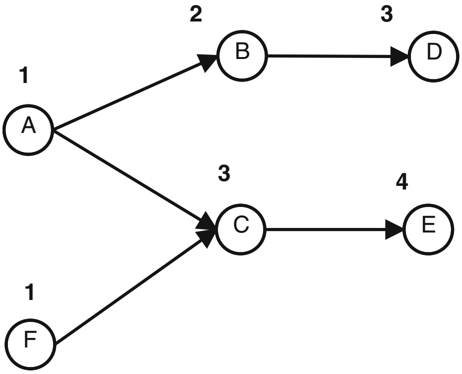
Article	1-gen citations	2-gen citations
C	2	0
E	1	2
D	1	1
B	1	0
A	0	0
F	0	0

where $f(A)$ is the f-value of article A, and A_i, $i = 1\ldots m$ are the articles citing article A. According to the equation, the minimum f-value for a published article is 1. Thus, the f-value of article A is 1 plus the sum of the f-values of all articles citing article A.

By performing the calculations for the articles of citation graph in Fig. 2, we produce the graph shown in Fig. 3, with the number on top of the nodes representing the f-values for the corresponding articles.

Such an approach results to each article eventually receiving thus much credit as the sum of the credit received by all articles that cite it, making no distinction between direct or indirect citations. This is also obvious by examining the results shown in Fig. 3. The f-value of each article is 1 plus the f-values of all direct citations. Of special interest are the f-values of articles C and D which are both 3. This means that based on Eq. 3 these two articles are equally important even though article C has received 2 1-gen citations and article D has received one 1-gen citation and one 2-gen citation.

Fig. 3 f-Values for Citation Graph 2



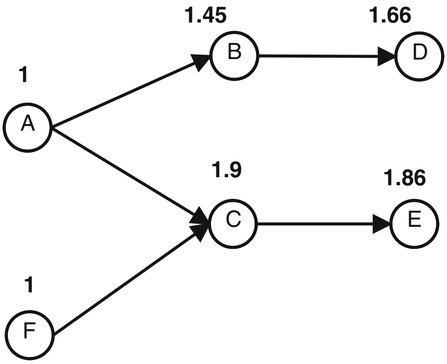


Fig. 4 f-Values for Citation Graph 2

So, there must be some factor that will assist us in differentiating direct and indirect citations. This is going to be a value that will reduce the cascaded f-value passed to an article's direct citations. Here is the new equation that calculates the f-value of an article:

$$f(A) = 1 + RF * (f(A_1) + f(A_2) + \cdots + f(A_m)) \tag{4}$$

For the dataset used in this paper we have calculated that $RF = 2.2$. The method for calculating it, is presented at "c^2-IF algorithm results and statistical analysis" section. Figure 4 demonstrates the use of $RF = 2.2$ on citation graph 2.

Determining the reducing factor

In this section we explain how the reducing factor (*RF*) is calculated. First, we provide a description of the CiteSeer database and the preprocessing we performed on it. Then, we use cc-IF information up to depth 3 to compute statistical information which we then use to calculate the reducing factor of the CiteSeer database.

Data used

We chose the CiteSeer database because:

- It indexes a sufficient number of research articles and is not limited to certain journals
- It mostly covers the scientific area of Computer and Information Science
- it uses the Open Access Initiative (OAI) format, which is XML based.

A sample record is shown in Fig. 5. For simplicity, only the identifiers that are used by the algorithm are listed.

Each article is defined by a unique <*identifier*> tag generated by CiteSeer, as shown in Fig. 5. Other fields required by the algorithm are the title (<*dc*:*title*> tag) and the list of references included in each article (<*oai_citeseer*:*relation*> tag).

Preprocessing

The original data consisted of the entire CiteSeer database; a total of 72 files, each holding 10,000 articles with their corresponding bibliographic details. Articles appearing in the list of references of a particular article are also part of the CiteSeer database. In order to

Fig. 5 CiteSeer Record

```
<record>
<header>
<identifier>oai:CiteSeerPSU:number#</identifier>
</header>
<metadata>
<dc:title>The Title</dc:title>
<oai_citeseer:pubyear>Publication Year</oai_citeseer:pubyear>
<oai_citeseer:relation type="References">
<oai_citeseer:uri>oai:CiteSeerPSU:number#</oai_citeseer:uri>
</oai_citeseer:relation>
<oai_citeseer:relation type="References">
<oai_citeseer:uri>oai:CiteSeerPSU:number#</oai_citeseer:uri>
</oai_citeseer:relation>
</oai_citeseer:oai_citeseer>
</metadata>
</record>
```

retrieve the necessary information and to store it in the relational database we developed a parsing algorithm.

During the parsing process certain errors occurred, mainly concerning articles with insufficient information. For the algorithms presented here, articles lacking information about their authors (26,040 in total) or their publication year (280,098 in total) where excluded from the procedure.

c^2-IF algorithm results and statistical analysis

The c^2-IF algorithm presented in (Fragkiadaki et al. 2009) calculates the numbers of direct and indirect citations present in a Citation Graph, up to a pre-specified depth (in this case up to depth 3). Moreover, it stores in the relational database all the paths in the citation graph that produce these citations thus giving us complete knowledge of the graph. We note that the database stores information about 410,205 articles, with 265,563 identified authors and 1,245,171 direct references among the articles.

During the processing of the data stored in the database we detected many cases where an article cites articles with future publication dates, for example, article A published in 1995 cites article B published in 2000. This situation creates cycles in the citation graph which lead to inaccurate results. In order to avoid such anomalies, we remove from the reference list of every citing article the articles published on the same year as the citing article or a future year. In other words, every article in the database is "allowed" to only cite articles published prior to itself. All other citations (arcs) are excluded from the original dataset. Thus, the direct references among articles in the database were reduced from 1,245,171 to 1,000,077.

After the execution of the algorithm, 1,000,077 1-gen citations, 4,095,493 2-gen citations and 14,924,150 3-gen citations were detected among the articles and that many paths were stored in the database. An interesting fact is that from the 410,025 articles originally included in the database only 133,658 receive at least one citation. To gain a better understanding of our data we calculated the summary statistics for each n-gen ($n = 1, 2, 3$) citation type (see Table 2).

If we compare the mean to the median we observe that in all three cases the median is lower than the mean. This means that even though the means are high they are mostly affected by a small number of articles with high values. This hypothesis is proven true if

Table 2 Summary statistics for 1-gen, 2-gen and 3-gen citations

	1-gen	2-gen	3-gen
Mean	7.48	30.64	111.7
SD	18.98	139.36	774.38
Min	1	0	0
25%	1	0	0
Median	3	2	0
75%	7	15	18
Max	1,280	12,186	82,182

we examine the quartile information. For example, for 1-gen citations we find that at least 75% of the articles in our database have fewer 1-gen citations than the corresponding mean value, whereas, the maximum value is 1,280 which is much larger than the usual values calculated for articles. Even greater are the differences for 2-gen citations and 3-gen citations.

Finally we identified the ratios

$$\frac{\text{number of 2-gen citations}}{\text{number of 1-gen citations}} \tag{5}$$

and

$$\frac{\text{number of 3-gen citations}}{\text{number of 2-gen citations}} \tag{6}$$

for all articles in our database and we calculated the corresponding summary statistics shown in Table 3.

We observe, that on average, for each 1-gen citation an article receives from within our database, it also receives 2.22 2-gen citations and for each 2-gen citation it receives 1.54 3-gen citations. This is an expected result since according to the definition of n-gen citations, the $(n+1)$-gen citations an article receives is the sum of all 1-gen citations received by the n-gen citations of the article. For example the 2-gen citations received by an article are the sum of all 1-gen citations received by the articles directly citing the article in question (1-gen citations). We also mention that there are 44,280 articles for which we can not calculate ratio 6 because the number of 2-gen citations they have received so far is 0.

Based on these statistical data we chose to use 1/2.2 as a reducing factor for the calculation of the f-value. We expect this value to differ among scientific areas or bibliographic databases.

Table 3 Summary statistics for the ratios in Eqs. 5 and 6

	2-gen/1-gen	3-gen/2-gen
Mean	2.22	1.54
SD	4.92	2.48
Min	0	0
25%	0	0
Median	1.00	0.91
75%	2.643	2.10
Max	454	227

f-Value algorithm

In this section we present the algorithm that calculates the f-values of all articles in our bibliographic database. This algorithm requires a finite number of iterations to calculate the f-values.

The algorithm receives as input the list of articles to be processed (I), the *Article Direct Citations* (ADC) data structure which includes for each article the list of articles that cite it, and, the *Article F-Values* (AFV) data structure which includes the articles that need to be processed plus their current f-value and a flag that denotes whether this value has changed since the last iteration. In other words, if we denote an article by R_x, then for a database with m articles, the list of all articles that need to be processed is $I = [R_1, R_2, R_3, \ldots, R_m]$. Let CR_x denote the list of articles that reference R_x. Thus, CR_x is a subset of I and the Article Direct Citations (ADC) data structure is $ADC = [CR_1, CR_2, CR_3, \ldots, CR_m]$. Additionally, for each article R_x, let VR_x denote the information required for this article during the execution of the algorithm. This information consists of the f-value calculated so far for this article and of a flag indicating whether the f-value has changed since the last iteration of the algorithm. Thus, VR_x= [fval = 1, changed = 0] for every article $R_x$in the beginning of the algorithm. Finally, the Article F Values structure is $AFV = [VR_1, VR_2, \ldots, VR_m]$. The algorithm returns the AFV structure with the calculated f-values for all articles in the database.

During the first iteration of the algorithm, all articles have an f-value equal to 1. At each iteration, the algorithm calculates the f-values of all articles in the database based on the f-values calculated during the previous iteration and records whether any f-value has changed between the two iterations. If there is at least one changed value, the algorithm requires one more iteration because that change could propagate to more articles in the following iteration. If there is no f-value change then all f-values have been calculated and the algorithm terminates.

Algorithm 1 f-Value algorithm

```
Input:
  I list of articles to be processed
  ADC data structure with direct citations of each article
  AFV data structure with initial f-values and flags
Output:
  AFV data structure with calculated f-values and flags

ADC = remove_cycles(ADC)
NChanged = 0
first = true
while (first || NChanged > 0) do
  first = false
  NChanged = 0
  PREV_AFV = AFV
  for each R in I do
    prev_fval = AFV[R][fval]
    AFV[R][fval] = 1
    RCIT = ADC[R]
```

Algorithm 1 continued

```
for T in RCIT do
  AFV[R][fval] = AFV[R][fval] + RF*PREV_AFV[T][fval]
if AFV[R][fval] != prev_fval then
  AFV[R][changed] = 1
  NChanged = NChanged + 1
else
  AFV[R][changed] = 0
```

In order to avoid possible errors in the execution of the algorithm we must ensure that no cycles exist in the collection of articles stored in our database. Since the algorithm calculates the f-value of an article based on the f-values of the articles that cite it, if there is a cycle the algorithm will enter an infinite loop.

Experimental results

In order to compare the three different indicators for measuring an article's scientific impact, we tested them against our database and report the obtained rankings per indicator. Recall that only 133,658 out of 410,025 articles listed in our database actually receive at least one 1-gen citation. In addition, there are 203,607 articles that do not give any citation, 38,100 of which receive citations from other articles while the rest do not give or receive any citations. Apart from presenting the rankings, the tables are complemented with the c^2-IF Information about the *n*-gen citations received by the articles up to depth 3. This information derives from the c^2-IF algorithm originally introduced at (Fragkiadaki et al. 2009). The algorithm was modified for the needs of the present paper. Table 4, shows the top 10 articles according to the received number of citations.

In order to test the Page Rank algorithm for citation graphs against our bibliographic database, we used an implementation written by Vincent Kräutler in Python (Kräutler 2006), which is based on a mathematical essay by Austin (2006). The implementation of the Page Rank algorithm as a package was imported to a Python script created for handling the reading/writing from/to the database and transforming the data into the appropriate format. The results are shown in Table 5.

Algorithm 1 was implemented and executed against our database. Table 6 shows information about the top 10 ranked articles.

Finally, Table 7 shows the summary statistics for all three approaches.

Discussion

In this section, we comment on the similarities and differences of the three indicators. In addition, we attempt to interpret the experimental results we obtained.

The Number of Citations, a measure used traditionally in citation analysis, plays an important role in all indicators. In Page Rank, the direct citations a publication receives are referred to as inbound links to its node in the citation graph and they are similarly used in the calculations of the f-value.

Table 4 Number of citations: top 10 ranked articles

Rank	Article title	Pub. year	Num. of citations	c^2-IF information		
				1-gen	2-gen	3-gen
1	Graph-Based Algorithms for Boolean Function Manipulation	1986	1,280	1,280	7,057	31,724
2	Optimization by Simulated Annealing	1983	1,027	1,027	4,508	17,090
3	Congestion Avoidance and Control	1988	879	879	12,186	92,182
4	A Method for Obtaining Digital Signatures and Public-Key Cryptosystems	1978	867	867	7,678	42,807
5	Statecharts: A Visual Formalism For Complex Systems	1987	803	803	3,590	12,045
6	Random Early Detection Gateways for Congestion Avoidance	1993	762	762	6,244	31,185
7	Fast Algorithms for Mining Association Rules	1994	735	735	3,681	12,688
8	Tcl and the Tk Toolkit	1994	700	700	4,726	22,976
9	Chord: A Scalable Peer-to-Peer Lookup Service for Internet Applications	2001	610	610	1,672	1,351
10	Mining Association Rules between Sets of Items in Large Databases	1993	594	594	5,178	22,961

Table 5 Page Rank: top 10 ranked articles

Rank	Article title	Pub. year	PR value	c^2-IF information		
				1-gen	2-gen	3-gen
1	Optimization by Simulated Annealing	1983	$686, 054*10^{-9}$	1,027	4,508	17,090
2	Graph-Based Algorithms for Boolean Function Manipulation	1986	$662, 149*10^{-9}$	1,280	7,057	31,724
3	New Directions in Cryptography	1976	$576, 792*10^{-9}$	422	5,224	34,203
4	A Method for Obtaining Digital Signatures and Public-Key Cryptosystems	1978	$526, 387*10^{-9}$	867	7,678	42,807
5	Congestion Avoidance and Control	1988	$461, 410*10^{-9}$	879	12,186	92,182
6	Applications Of Circumscription To Formalizing Common Sense Knowledge	1986	$323, 209*10^{-9}$	226	2,611	16,881
7	Tcl and the Tk Toolkit	1994	$315, 861*10^{-9}$	700	4,726	22,976
8	Implementing Mathematics with The Nuprl Proof Development System	1986	$309, 963*10^{-9}$	398	3,858	17,598
9	Statecharts: A Visual Formalism For Complex Systems	1987	$309, 718*10^{-9}$	803	3,590	12,045
10	A Timeout-Based Congestion Control Scheme for Window Flow-Controlled Networks	1986	$304, 607*10^{-9}$	26	1,486	21,789

In general, the latter two approaches are based on the assumption that the use of the Number of Citations as a measurement of the importance of a scientific publication is insufficient. The resulting ranking is solely based on the direct impact the article has without taking into account its present state (whether it remains in the researchers'

Table 6 f-Value: top 10 ranked articles

Rank	Article title	Pub. year	f-Value	c^2-IF information		
				1-gen	2-gen	3-gen
1	Congestion Avoidance and Control	1988	258,534	879	12,186	92,182
2	Design and Implementation of the Sun Network Filesystem	1985	234,037	296	4,299	39,239
3	The UNIX Time-Sharing System	1974	224,167	127	1,405	14,236
4	A Scheme for Real-Time Channel Establishment in Wide-Area Networks	1990	192,736	421	5,172	46,302
5	A Timeout-Based Congestion Control Scheme for Window Flow-Controlled Networks	1986	181,751	26	1,486	21,789
6	A Fast File System for UNIX	1984	148,843	83	1,610	13,429
7	New Directions in Cryptography	1976	138,137	422	5,224	34,203
8	An Open Operating System for a Single-User Machine	1979	114,979	12	878	9,894
9	Using Sparse Capabilities in a Distributed Operating System	1986	109,455	51	523	5,418
10	Why Aren't Operating Systems Getting Faster As Fast As Hardware?	1989	103,480	149	2,451	19,929

Table 7 Summary statistics

	Number of citations	Page Rank	f-Value
Mean	7.48	$2, 451*10^{-9}$	43.06
SD	18.98	$4, 258*10^{-9}$	1,221
Min	1	$1, 788*10^{-9}$	1
25%	1	$1, 788*10^{-9}$	1
Median	3	$1, 788*10^{-9}$	1
75%	7	$2, 011*10^{-9}$	1.66
Max	1,280	$686, 954*10^{-9}$	258,534

preferences) or its derived contribution (the impact it has on the research in the specific scientific field). The f-value indicator and Page Rank appear to be very similar in nature, thus, before elaborating on their experimental results, we discuss their main differences and similarities. These are summarized in the following:

1. *The logic behind the equation:* Page Rank focuses on a person (the "random scientist") moving from article to article randomly by choosing to read next an article that appears as a citation in the List of References of the article she reads. All cited articles have the same probability to be selected. The f-value is not based on such a probability, but on the cumulative value of the *n*-gen citations that an article has received.
2. *How are citations treated:* Page Rank for Citation graphs divides equally the value of an article among its cited articles. Such a division implies that among two articles with equal values, A and B, if A cites 10 articles and B cites 20 articles, then articles cited by A will receive twice as much recognition than articles cited by B, just because A has cited fewer articles. Since we cannot assume that cited articles have less impact

when they are encountered in longer reference lists, we claim that this division of value does not correspond to a real world behavior, thus, it is not included in the calculations of an article's f-value.

3. *The damping factor:*In the f-value calculation there is no damping factor. Instead, there is a reducing factor used to dicrease the accumulated value of the *n*-gen citations. This factor has been chosen to be $\frac{1}{2.2}$ (see "Determining the reducing factor" section). In addition, the f-value also has a minimum value of 1 for all articles. The f-value of an article always increases as more articles cite directly and/or indirectly the article in question.

Even though the equations used in the calculation of the Page Rank for Citation Analysis and the f-value appear similar, the logic behind each approach is differenet.

We now proceed and discuss the experimental results in an effort to better understand the differences and similarities among the three indicators. Examining the top 10 ranked articles based on the Number of Citations (Table 4), it is very interesting to notice the c^2-IF information provided, especially for the top four ranked articles. We observe that according to this indicator, the "Congestion Avoidance and Control" article is ranked 3rd, because it has received fewer direct citations than the two articles above it. On the other hand, if we examine the c^2-IF information, we can clearly see that it has received considerably more 2-gen citations and 3-gen citations than the first and second ranked articles. The same is true to a lesser extent for the fourth ranked article. But, this information is not taken under consideration for this ranking.

Table 5, shows the top 10 articles based on PageRank along with the corresponding c^2-IF Information. The ranking is different here, and, by inspecting the c^2-IF information of the top two articles, we observe that the first ranked article has less 1-gen, 2-gen and even 3-gen citations than the second ranked article. This ordering can only be explained if we consider the way Page Rank values are calculated. Apparently, the "Optimization by Simulated Annealing" article has received fewer 1-gen, 2-gen and 3-gen citations than the second article as an absolute number, but, the prestige (Page Rank value) of the articles that cite it played an important role in the calculations. In addition, the number of citations made by the citing articles has also affected the result. So, we have to assume that although the up to 3-gen citations of the first article are fewer than the ones received by the "Graph-Based Algorithms for Boolean Function Manipulation" article, they are either of higher value and/or have a smaller number of outbound links.

The f-value results are presented in Table 6 along with the corresponding c^2-IF information. Let us examine the first ranked article. This article was ranked third according to the Number of Citations. This is explained by the fact that the calculation of the f-value is exhaustive in nature and takes into consideration all the knowledge present in the citation graph. In other words, an article's f-value increases as it receives more citations at each depth, all the way to the longest citation path.

Finally, Table 8 shows all articles listed in Tables 4, 5 and 6 along with their c^2-IF information. The articles are ordered by their f-value rank. Again, we observe that the rankings vary significantly depending on the indicator used.

The first approach, Number of Citations, only takes into account the direct impact an article has based on the number of citations it receives. On the other hand, Page Rank does not take into account the direct impact alone but it also considers, to some extent, the added value provided by the citing articles of the article in question. We should point out though that Page Rank is not an exhaustive method, that is, for the calculation of the importance

Table 8 Summarized results of Top article rankings based on all three approaches

Article title	Pub. year	Ranks			c^2-IF information		
		f-value	Number of citations	Page Rank	1-gen	2-gen	3-gen
Congestion Avoidance and Control	1988	1	3	5	879	12,186	92,182
Design and Implementation of the Sun Network Filesystem	1985	2	75	20	296	4,299	39,239
The UNIX Time-Sharing System	1974	3	498	39	127	1,405	14,236
A Scheme for Real-Time Channel Establishment in Wide-Area Networks	1990	4	26	11	421	5,172	46,302
A Timeout-Based Congestion Control Scheme for Window Flow-Controlled Networks	1986	5	7,365	10	26	1,486	21,789
A Fast File System for UNIX	1984	6	1,126	139	83	1,610	13,429
New Directions in Cryptography	1976	7	23	3	422	5,224	34,203
An Open Operating System for a Single-User Machine	1979	8	18,272	268	12	878	9,894
Using Sparse Capabilities in a Distributed Operating System	1986	9	2,608	323	51	523	5,418
Why Aren't Operating Systems Getting Faster As Fast As Hardware?	1989	10	365	182	149	2,451	19,929
A Method for Obtaining Digital Signatures and Public-Key Cryptosystems	1978	19	4	4	867	7,678	42,807
Applications Of Circumscription To Formalizing Common Sense Knowledge	1986	71	143	6	226	2,611	16,881
Graph-Based Algorithms for Boolean Function Manipulation	1986	76	1	2	1,280	7,057	31,724
Random Early Detection Gateways for Congestion Avoidance	1993	129	6	34	762	6,244	31,185
Tcl and the Tk Toolkit	1994	131	8	7	700	4,726	22,976
Implementing Mathematics with The Nuprl Proof Development System	1986	156	35	8	398	3,858	17,598
Optimization by Simulated Annealing	1983	168	2	1	1,027	4,508	17,090
Mining Association Rules between Sets of Items in Large Databases	1993	249	10	23	594	5,178	22,961
Statecharts: A Visual Formalism For Complex Systems	1987	326	5	9	803	3,590	12,045
Fast Algorithms for Mining Association Rules	1994	531	7	25	735	3,681	12,688
Chord: A Scalable Peer-to-Peer Lookup Service for Internet Applications	2001	2,621	9	150	610	1,672	1,351

of a research article one does not traverse the entire citation graph. Finally, in the calculations of the f-value, the indirect impact an article has is fully accumulated in the calculations. The whole citation graph is traversed and the value of each article is partially propagated to all articles that it cites, thus producing an exhaustive method that uses all the information present in the citation graph.

The calcualtions for the f-value indicator are based on historical data, that is, they are dependent on the dataset. It is very likely that the reducing factor will be different for different datasets. A different reducing factor is expected to alter the resulting ranking, but the extend at which the ranking is affected requires more research.

Conclusions

Based on the Cascading Citations Indexing Framework, we proposed a new indicator for measuring the importance of a research article. The f-value represents a unique value for each article that takes into consideration the *n*-gen citations received by the specific article. We developed an algorithm that calculates the f-value for all articles in a bibliographic database, and we experimentaly compared it to two other indicators.

Future work on this field will: (a) try to incorporate other aspects of the c^2-IF in the calculation of the f-value, (b) examine the impact the different values of the reducing factor have on the final ranking of the articles, and, (c) examine whether there can be a unified f-value for interdisciplinary articles.

References

Anderson, T.R., Hankin, R. K. S., & Killworth P. D. (2008) Beyond the durfee square: Enhancing the h-index to score total publication output. *Scientometrics 76*(3), 577–588. doi:10.1007/s11192-007-2071-2.

Austin, D. (2006). How Google finds your needle in the web's haystack. http://www.ams.org/featurecolumn/archive/pagerank.html.

Boldi, P., Santini, M., & Vigna, S. (2009). Pagerank: Functional dependencies. *ACM Transactions on Information Systems 27*(4), 1–23. doi:10.1145/1629096.1629097.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems, 30*, 107–117.

Citeseer. (1997). http://www.citeseer.ist.psu.edu.

Dervos, D., & Kalkanis, T. (2005). cc-IFF: A cascading citations impact factor framework for the automatic rankings of research publications. In *3rd IEEE international workshop on intelligent data acquisition and advanced computer systems: technology and applications (IDAACS 2005), Sofia, Bulgaria.*

Dervos, D., & Klimis, L. (2008). Exploiting cascading citations for retrieval. In *Proceeding of the ASSIST 2008 annual meeting*.

Dervos, D., Samaras, N., Evangelidis, G., & Folias, T. (2006). A new framework for the citation indexing paradigm. In *Proceedings of the ASSIST 2006 annual meeting, Austin, Texas, USA*.

Egghe, L. (2006). Theory and practise of the g-index. *Scientometrics, 69*(1), 131–152.

Fragkiadaki, E., Evangelidis, G., Samaras, N., & Dervos, D. (2009). Cascading citations indexing framework algorithm implementation and testing. *Informatics, Panhellenic conference on Informatics*, 70–74. doi:10.1109/PCI.2009.30.

Garfield, E. (1955). Citation indexes for science. A new dimension in documentation through association of ideas. *Science 122*, 1123–1127.

Garfield, E. (1999). Journal impact factor: A brief review. *CMAJ 161*(8), 979–980.

Garfield, E. (2005). The agony and the ecstasy—the history and meaning of the journal impact factor. In *International Congress on Peer Review And Biomedical Publication*.

Giles, C. L., Bollacker, K. D., & Lawrence, S. (1998). *Citeseer: An automatic citation indexing system* (pp. 89–98). New York: ACM Press.

Guns, R., & Rousseau, R. (2009). Real and rational variants of the h-index and the g-index. *Journal of Informetrics, 3*(1): 64–71. doi:10.1016/j.joi.2008.11.004.

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

Jin, B., Liang, L., Rousseau, R., & Egghe, L. (2007). The R- and AR-indices: Complementing the h-index. *Chinese Science Bulletin, 52*(6), 855–863. doi:10.1007/s11434-007-0145-9.

Katsaros, D., Sidiropoulos, A., & Manopoulos, Y. (2007). Age decaying h-index for social network of citations. In *SAW proceedings of the BIS 2007 workshop on social aspects of the web, Poznan, Poland*, April 27, 2007, CEUR-WS.org. *CEUR workshop proceedings*. 245.

Kräutler, V. (2006). The Google pagerank algorithm in 126 lines of Python. http://www.kraeutler.net/vincent/essays/googlepagerankinpython.

Ma, N., Guan, J., & Zhao, Y. (2008). Bringing pagerank to the citation analysis. *Information Processing and Management, 44*(2): 800–810. doi:10.1016/j.ipm.2007.06.006.

Rousseau, R. (1987). The Gozinto theorem: Using citations to determine influences on a scientific publication. *Scientometrics, 11*(3–4): 217–229.

Sidiropoulos, A., Katsaros, D., & Manolopoulos, Y. (2007). Generalized hirsch h-index for disclosing latent facts in citation networks. *Scientometrics, 72*(2), 253–280 doi:10.1007/s11192-007-1722-z.