Using Hypertext Metrics to measure Research Output Levels

Paul De Bra[*]
Department of Computing Science
Eindhoven University of Technology
PO Box 513
NL 5600 MB Eindhoven
The Netherlands
debra@win.tue.nl

July 21, 1999

Abstract

Two common ways to measure the "output" of a researcher (or research group) are to count numbers of publications and to count the citations (references to these publications in publications of others). These simple methods are flawed because they cannot easily take into account the differences in publication and citation habits in different scientific communities.

An alternative approach is to view citations as hypertext links, and to use or adapt hypertext metrics to compare the scientific output of researchers, in comparison to that of others in areas with similar publication and citation patterns. We show how hypertext metrics, introduced by Botafogo, Rivlin and Shneiderman, can be modified in order to identify comparable research fields based on their publication and citation pattern. An author's performance can then be compared to that of others in research fields with a similar pattern.

[*]Paul De Bra is also affiliated with the University of Antwerp, Belgium, and with the "Centrum voor Wiskunde en Informatica (CWI) in Amsterdam.

1 Introduction

Measuring research output "accurately" and comparing different researchers' output "fairly" are probably unattainable goals. Some say that having many publications is most important, while others say that being cited by others is a more important sign of research quality. Furthermore there is no consensus on which publications (or citations from publications) should count, based on how important the journals or conference proceedings are in which articles are published.

This paper does not enter the discussion on the dismissal of publications and citations based on journal or conference. It does attempt to provide a way to decide "how well" a researcher has been publishing, compared to others in fields with a similar publication and citation pattern. Such comparable research field may deal with completely different topics, while closely related research fields may not have a comparable publication and citation pattern at all. Some previous work on the definition of metrics for citations [5, 4] tries to come up with absolute measures of a researcher's track record without considering the "average" publication and citation behavior in the author's research field.

In order to determine what constitutes a "research field" one may use citation patterns as well. In [2] an algorithm is given for determining "Communities" on the Web. The same algorithm can deduce research fields from a citation database. In this paper we assume that such an algorithm was used to construct a "citation database" for a specific research field.

In [1] an attempt is made to characterize hypertext link structures. The current paper reuses some metrics from that paper for analyzing publication and citation behavior. Section 2 recalls the metrics from [1]. Section 3 adapts these metrics to citation link structures and derives some new metrics. Section 4 concludes and shows possible future work.

2 Hypertext Metrics Revisited

In [1] several aspects of a hypertext document's link structure are studied, in order to compare different documents and to find potential problem spots. In [3] this study is carried even further to also include the user's navigation strategy. Three important observations in the use of hypertext lead to three hypertext metrics:

- When there are few links in a hyperdocument, pages (nodes) are difficult to reach by following links. Users may become disoriented because they need to go through many steps to get to some piece of information they want. Some parts of a hyperdocument may not even be connected by links at all. When there are a lot of links, the path to a specific node will (on average) be short, but the abundance of links implies that the links do not help the user in selecting interesting places to go. *Compactness* is a metric which tries to capture how well connected (by links) a hyperdocument is.
- When looking at the hierarchy, by following links from a given starting point (node), along the shortest path down to each individual node, these paths can be short in

the case of a *shallow* and *wide* hierarchy or long in the case of a *deep* and *narrow* hierarchy. The term *stratum*, taken from the study of organizations, is used for a metric of how deep or linear the link structure is, and thus of how much of a reading (or navigation) order is imposed on the user.

- Apart from the links that are used in determining the hierarchical link structure, a hyperdocument may also contain links that represent shortcuts to navigate between (arbitrary) nodes. These are called *cross-reference* links. The *ratio of hierarchical vs. cross-reference links* can be used as a metric, with similar intentions as compactness: if there are few cross-reference links it takes many steps to navigate between two arbitrary nodes, and if there are many cross-reference links the link structure becomes a dense mess in which links offer no guidance.

When we translate these observations to literature references (in scientific papers) instead of hypertext link structures, we can easily give their corresponding meaning:

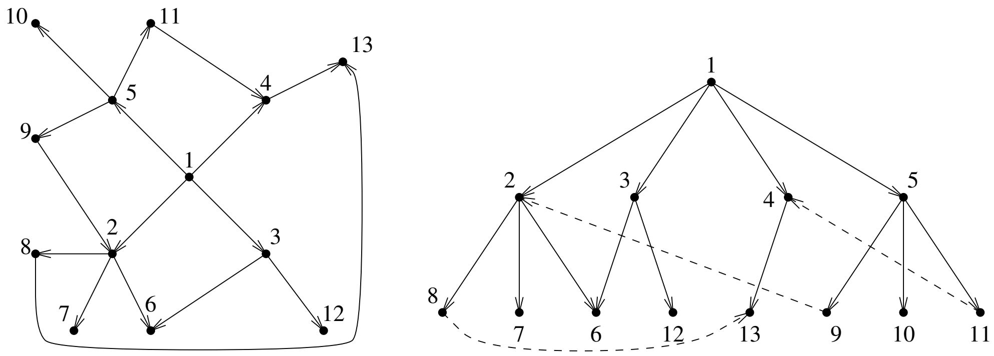
- When we're studying the literature in a certain field we may find large differences in the density of citations. (The number of references in each paper is always very small in some research fields, whereas it is very large in other fields.) We may also find clusters that are strongly connected, whereas there are little or no references to another cluster. Compactness is a rough estimate of the citation behavior of a research community as a whole. It can be used to identify other research fields with a similar citation behavior (similar compactness).
- Stratum measures the "relative depth" of a link structure. In terms of citations it indicates how citations to an article are spread over time. Are there many citations in a short period of time or do citations spread out over a longer period? In some research fields old papers are still cited directly whereas in other fields there are only references to recent literature.
- There are actually two types of cross-reference links: ones that merely form shortcuts and ones that create cycles in the link structure. Citation databases are typically acyclic. Unless authors are aware of each others' upcoming publications they do not create references to future papers. For this reason we will not discuss the ratio of hierarchical versus cross-reference "links", and explicitly use the acyclic nature of the link structure in citations.

2.1 The (Converted) Distance Matrix

In order to calculate metrics on distances between hypertext nodes (measured in number of links to follow) the following matrix is used:

Definition 2.1 *Given a hyperdocument H with N nodes, the distance matrix D is an $N \times N$ matrix such that for every pair of nodes i and j :*

- $D[i,j]$ *is the length of the shortest path (in number of forward links) from* i *to* j*, if such a path exists;*
- $D[i,j]$ *is* ∞ *otherwise.*



"Graph representation" "Tree approximation"

	1	2	3	4	5	6	7	8	9	10	11	12	13
1	0	1	1	1	1	2	2	2	2	2	2	2	2
2	∞	0	∞	∞	∞	1	1	1	∞	∞	∞	∞	2
3	∞	∞	0	∞	∞	1	∞	∞	∞	∞	∞	1	∞
4	∞	∞	∞	0	∞	∞	∞	∞	∞	∞	∞	∞	1
5	∞	∞	∞	∞	0	∞	∞	∞	1	1	1	∞	∞
6	∞	∞	∞	∞	∞	0	∞	∞	∞	∞	∞	∞	∞
7	∞	∞	∞	∞	∞	∞	0	∞	∞	∞	∞	∞	∞
8	∞	∞	∞	∞	∞	∞	∞	0	∞	∞	∞	∞	1
9	∞	1	∞	∞	∞	2	2	2	0	∞	∞	∞	3
10	∞	∞	∞	∞	∞	∞	∞	∞	∞	0	∞	∞	∞
11	∞	∞	∞	1	∞	∞	∞	∞	∞	∞	0	∞	2
12	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	0	∞
13	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	0

Figure 1: Example link structure and distance matrix.

Figure 1 (taken from an example in [1]) shows a small link structure and the corresponding distance matrix. (Note that in the matrix the vertical dimension is the first coordinate and the horizontal the second, as was done in [1].) We cannot perform simple calculations like "take the average of the values in the distance matrix" because of the occurrence of ∞. In order to obtain meaningful (or at least finite) results, ∞ can be replaced by a large value, called the *conversion constant*. To avoid interference between this constant and the existing finite values in the distance matrix, the conversion constant needs to be larger than the maximum length of the shortest path between two nodes (measured in forward links). For this reason we usually take the number of nodes, N, as the conversion constant.

Definition 2.2 *The* converted distance matrix *for a hyperdocument with N nodes is obtained by replacing each occurrence of ∞ in the* distance matrix *for that hyperdocument by N.*

2.2 Compactness

The *compactness* metric tries to measure the average length of the shortest path between two arbitrary nodes, or in other words how well connected the link structure is. In order to obtain values that are comparable for small and large hyperdocuments the metric is defined in a normalized way (resulting in values between 0 and 1).

Definition 2.3 *Given a hyperdocument H with N nodes and converted distance matrix D, the* compactness *of H is:*

$$\frac{Max - \sum_{i=1}^{N} \sum_{j=1}^{N} D[i,j]}{Max - Min}$$

where:

- $Max = N^3 - N^2$
- $Min = N^2 - N$

$Max = N^3 - N^2$ is the maximum value the sum of all elements of the converted distance matrix can assume, and is reached when the hyperdocument contains no links at all. $Min = N^2 - N$ is the minimum possible value of that sum, and is reached when the links form a complete graph (i. e. there are links in both directions between each pair of nodes).

A hyperdocument without links has a compactness of 0 and a complete graph has a compactness of 1. The small link structure of Figure 1 has a compactness of 0.18, and thus seems not as well connected. (In [1] the compactness of some well-known hypertext documents is found to be around 0.5.)

2.3 Stratum

Stratum measures how "linear" a hyperdocument is, or how "narrow" the link hierarchy, or in other words, how much the user is forced to navigate in a certain (linear) way.

Definition 2.4 *Let H be a hyperdocument with N nodes and distance matrix D. Let D' be derived from D by replacing all occurrences of ∞ by 0.*

The stratum *of H is*

$$\frac{4\sum_{i=1}^{N}\left(|\sum_{j=1}^{N}D'[j,i]-\sum_{j=1}^{N}D'[i,j]|\right)}{N^3}$$

if N is even and is

$$\frac{4\sum_{i=1}^{N}\left(|\sum_{j=1}^{N}D'[j,i]-\sum_{j=1}^{N}D'[i,j]|\right)}{N^3-N}$$

if N is odd.

In [1] it is explained that this metric consists of sums of (absolute values of) sums of distances from a node to other nodes minus distances from other nodes to that node. These numbers are also called the *absolute prestige* of the nodes, and their sum is the absolute prestige of the whole hyperdocument. The normalization factor $N^3/4$ or $(N^3-N)/4$ represents the absolute prestige of a linear document, also called the *linear absolute prestige* or LAP, which is maximal. (These formulas are proven in [1].) The stratum of a hyperdocument is a value between 0 and 1. The value 0 is the stratum of a circular link structure, and 1 is the stratum of a linear structure.

The stratum of the link structure of Figure 1 is 0.12. This fairly low value illustrates that there is a relatively high navigational freedom in this graph.

3 Publication and Citation Metrics

Although citation link structures resemble hypertext link structures we must adapt the definition of metrics in order to obtain more meaningful values.

3.1 Citation Compactness

Because of the acyclic nature of citation link structures the distance matrix will contain many occurrences of the value ∞. In fact, whenever $D[i,j]$ is not ∞, $D[j,i]$ is ∞. The result is that the occurrences of ∞ (or the conversion constant) dominate the calculation of the compactness. To avoid this problem we consider citations as bidirectional links for the calculation of compactness. (This results in $D[i,j]=D[j,i]$.)

Definition 3.1 *Given a set of literature references H with N papers and a converted distance matrix in which references and reverse references are taken into account. The* citation compactness *of H is:*

$$\frac{Max-\sum_{i=1}^{N}\sum_{j=1}^{N}D[i,j]}{Max-Min}$$

where:

- $Max = N^3 - N^2$
- $Min = N^2 - N$

As with *compactness* for hyperdocuments the *citation compactness* of a set of literature references is a number between 0 and 1.

Note that citation compactness shares an important property with compactness: it only considers *shortest paths*. Thus, when paper A has a reference to B, other indirect references like from A to C and from C to B have no influence on how the relationship between A and B is measured.

3.2 A Metric related to Stratum: Citation Prestige

The definition of stratum is based on the sum of the *prestige* of nodes. This term, taken from the study of organizations, weighs indirect relationships heavier than direct relationships. (The prestige of being someone's bosses boss is higher than just being that person's boss.) In citations, direct citations are more important than indirect references. To determine an *author's prestige* we take sum the prestige-values of all the author's papers.

Definition 3.2 *Let H be a citation database with N papers and distance matrix D. Let D' be defined as follows:*
$D'[i,i] = 0$ for all i. If $i \neq j$ then $D'[i,j] = 1/D[i,j]$ if $D[i,j]$ is finite.
$D'[i,j] = -1/D[j,i]$ if $D[i,j] = \infty$.

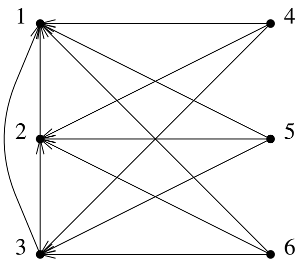
The citation prestige *of paper i is $\sum_{j=1}^{N} D'[j,i]$*

The citation prestige *of an author is the sum of the citation prestiges of all papers written by that author.*

Citation prestige is not independent of the size of the citation database. In order to normalize the citation prestige we divide by the maximum value an author can obtain in a citation database with N papers. Figure 2 shows a graph with 6 nodes. When nodes 1, 2 and 3 are written by the same author, that author has the maximum prestige. Possible links between papers 4, 5 and 6 have no influence on the citation prestige of the author of 1, 2 and 3. In this example the citation prestige of this author is $N-1 + N-3 + N-5$. Since $N = 6$ the prestige is actually 9 (and the *relative citation prestige* is 1). If papers 4, 5 and 6 were all written by one author the prestige of that author would be -9 (and the *relative citation prestige* would be -1).

Definition 3.3 *Let H, N, D and D' be as in Definition 3.2. The* relative citation prestige *of an author is the* citation prestige *of the author, divided by*

$$Max = \frac{N^2 - 3N}{2} \text{ when } N \text{ is even}$$



	1	2	3	4	5	6
1	0	-1	-1	-1	-1	-1
2	1	0	-1	-1	-1	-1
3	1	1	0	-1	-1	-1
4	1	1	1	0	0	0
5	1	1	1	0	0	0
6	1	1	1	0	0	0

Figure 2: Example citation structure needed for maximum author prestige.

or

$$Max = \frac{N^2 - 4N + 3}{2} \textit{ when } N \textit{ is odd.}$$

The prestige of an author is a value between -1 and 1 (although it is very unlikely that the extremes -1 and 1 are ever reached in practical cases).

3.3 Application of these Metrics

The performance of a researcher can be compared to that of others in *similar* research fields. We consider two research fields to be similar if their *citation compactness* and their *stratum* are “roughly” the same. Experience will tell how similar research fields need to be to be considered equivalent. We expect that deviations of 10% in compactness and stratum will be acceptable.

Within similar research fields the *relative citation prestige* of authors can be used as a measure of their performance. We argue that *relative citation prestige* is a usable metric because:

- The prestige increases as more papers cite the author’s papers. Having just one prestigious paper that is cited by everyone does not yield the maximum prestige. People must cite more papers by an author.

- The prestige decreases as the author writes more papers that are not cited by other papers.
- If we consider the publications of one author only, the relative citation prestige of the link structure of that author's self-citations is always 0. Thus, citing one's own papers does not add to one's prestige, but does not diminish it either.
- Groups of authors who frequently cite each other's work do not add to each other's prestige, but do not diminish it either.
- It is better to have a few important publications (often cited) than having many unimportant publications (seldom cited by others). Indeed, the references contained in the many unimportant papers actually decrease ones prestige.

4 Conclusions and Future Work

In this paper we have revisited some popular metrics for hypertext link structures, and proposed a way to reuse them in the area of measuring scientific output. We have defined a *relative citation prestige* metric that can be used to compare research output in similar research fields. (We also defined "similar".) Our *relative citation prestige* metric has the desirable property that it penalizes authors for publishing papers that are subsequently ignored by their research community and that it dismisses the "fake" prestige of small groups of authors who frequently quote each other.

Further research is needed to determine whether the metrics we defined are sufficiently independent of the size of research communities. Also, we wish to investigate how sensitive *relative citation prestige* is to variations in *citation compactness* and *stratum*.

References

[1] R. A. Botafogo, E. Rivlin, B. Shneiderman. Structural Analysis of Hypertexts: Identifying Hierarchies and Useful Metrics. ACM Transactions on Information Systems, 10:2, pp. 142–180, 1992.

[2] D. Gibson, J. Kleinberg, P. Raghavan. Inferring Web Communities from Link Topology. Proceedings of the Ninth ACM Conference on Hypertext and Hypermedia, pp. 225–234, Pittsburgh, 1998.

[3] J. De Vocht, Experiments for the Characterization of Hypertext Structures. Masters Thesis, Eindhoven Univ. of Technology, 1994.

[4] R. Philips, Pseudo-Metrics on Bibliographic Entities. Scientometrics, Vol. 25, nr. 1, pp. 193–199, 1992.

[5] R. Rousseau, The Gozinto Theorem: Using Citations to Determine Influences on a Scientific Publication. Scientometrics, Vol. 11, nr. 3–4, pp. 217–229, 1987.