PERGAMON

Information Processing and Management 38 (2002) 111–140

INFORMATION PROCESSING & MANAGEMENT

www.elsevier.com/locate/infoproman

Discovering authorities and hubs in different topological web graph structures

George Meghabghab *

Department of Computer Science Technology, Roane State, Oak Ridge, TN 37830, USA

Received 29 September 2000; accepted 17 January 2001

Abstract

This research is a part of ongoing study to better understand citation analysis on the Web. It builds on Kleinberg's research (J. Kleinberg, R. Kumar, P. Raghavan, P. Rajagopalan, A. Tomkins, Invited survey at the International Conference on Combinatorics and Computing, 1999) that hyperlinks between web pages constitute a web graph structure and tries to classify different web graphs in the new coordinate space: out-degree, in-degree. The out-degree coordinate is defined as the number of outgoing web pages from a given web page. The in-degree coordinate is the number of web pages that point to a given web page. In this new coordinate space a metric is built to classify how close or far are different web graphs. Kleinberg's web algorithm (J. Kleinberg, Proceedings of the ACM-SIAM Symposium on Discrete Algorithms, 1998, pp. 668–677) on discovering ''hub web pages'' and ''authorities web pages'' is applied in this new coordinate space. Some very uncommon phenomenon has been discovered and new interesting results interpreted. This study does not look at enhancing web retrieval by adding context information. It only considers web hyperlinks as a source to analyze citations on the web. The author believes that understanding the underlying web page as a graph will help design better web algorithms, enhance retrieval and web performance, and recommends using graphs as a part of visual aid for search engine designers. © 2001 Elsevier Science Ltd. All rights reserved.

Keywords: Web algorithms; Web graph; Graph theory; Citation analysis; In-degree graphs; Out-degree graphs; Complete bipartite graphs; Bipartite graphs; General graphs; Linear algebra; Principal eigenvector; Principal eigenvalue; Hub web page; Authority web page; Web page as a hub web page and an authority web page

* Present address: 9134 Harlaxton Court, Knoxville, TN 37923, USA.
E-mail address: gmeghab@hotmail.com (G. Meghabghab).

0306-4573/01/$ - see front matter © 2001 Elsevier Science Ltd. All rights reserved.
PII: S0306-4573(01)00019-X

1. Introduction

Scientific citations have been studied for a long time. Citation analysis in the field of bibliometrics (Egghe & Rousseau, 1990) is the science of studying citations, their structure, and the evolution of a specific domain of knowledge from its citations. Many information sciences journals, e.g., JASIS (Small, 1973, 1986) have devoted issues and complete volumes to the exploration of such a field of knowledge. A citation is static and unidirectional. Once an article is published, no new references can be added to it. A citation can be used to evaluate its impact and influence over the whole body of knowledge in a given field. A citation fulfilling such a role in a given field becomes an authority in that field. Citations in a given journal follow the same principles, standards, and forms. Thus, the standard deviation of the content of two articles in a given journal is usually small. Human judgment of the technicality of an article keeps the quality of publication at high level although the noise is present, and thus a citation is more relevant and more objective to a given topic. Citations link articles that are relevant to a given research. Garfield's impact factor (Garfield, 1972) is the most important factor ever developed to assess a journal j influence over the publication in that field. The impact factor is the average number of citations a journal j in a given year receives from other journal articles after it has been published for the last two years. It becomes the in-degrees of nodes in a given network of publications in that field. Pinski and Narin (1976) argued that a journal is influential if it is heavily cited by other influential journals. Citations of a given publication are "signatures of intelligence" of the space and time of a collective group of people.

A web page is dynamic and bi-directional. It points to other web links as well as other web links point to it. A web page gets updated and so new web links are discovered as new links are added. It is a living community of links. The contextual amount of variability between two web pages could be very high. Not only they deal with different subjects, but they could be in different languages, have nothing in common, no standards are imposed on the content, and reflect subjective ideas than commonly established scientific explorations. Human judgment on web content is more subjective, and noisier than in a citation. It is hard to keep a control on the quality of web pages that are created. Web pages serve many purposes not just to link two or more web pages. Hubs and authorities (Kleinberg, 1998) have been developed to better assess the influence that a web page has on other web pages that link to it. Web pages also reflect the signatures of intelligence in our era and contain rich information on our collective society as a whole.

Citation analysis is a very important field for information scientists for mining citations in journals. The comparison above sheds some light on the parallel between a web page and a citation in a journal (Kleinberg, Kumar, Raghavan, Rajagopalan, & Tomkins, 1999). This just to emphasize the interest that web mining experts ought to have to analyze web links even though it is a task in its infancy, and even though that the principles that govern web pages and web links are different from those of scientific citations. Web pages are not scrutinized like scientific journals are. New links in a web page can have an impact on the whole search done and the always changing size of the web can reflect patterns of indexing, crawling for search engine designers that feed on these web links (Meghabghab, 2000). More importantly, users need to be aware not only of the set of results returned, but also of the set of results not returned or the percentage of "rejected web pages" for each query (Meghabghab, 2001a). A formula on how to measure the

goodness of a search engine (SE) emerges as follows from this study on the set of queries (SQ) used

$$G = (\mathrm{SE};\ \mathrm{SQ}) = \alpha * (\mathrm{Coverage}) + \beta * (\mathrm{Age}), \tag{0}$$

where Coverage is the coverage with the estimated web size at the time of the experiment, and Age is the "median age" of new document in the result set, α and β are constants each between 0 and 1. A visualization graph of the structure of the returned set of results and the rejected set of web pages will prove helpful for both users and search engine designers (Kopetzky & Kepler, 1999).

To fully comprehend and assess the discovery of hubs and authorities and the influence that hyperlinks between different web pages have over other web pages, graph theory can be used to better understand the measures developed in (Kleinberg, 1998). Graph theory (Kleinberg et al., 1999) has already been applied to the web. Nodes represent static html pages and hyperlinks represent directed edges. But never an attempt have been made to study web graphs in the (out-degree, in-degree) coordinate space, neither has citation analysis on the web has been applied in this new coordinate space and has revealed the complexity of citation analysis in a variety of web graphs. Graph theory can be used also to discover new patterns that appear in a citation graph. The same idea can be used to measure the distance between two web pages. Measuring the topological structure richness of a collection of web pages is an important aspect of web pages never explored before and never applied before on hubs and authorities.

The next section is a reminder on concepts borrowed from graph theory to help analyze hyperlinks, and the richness of the WWW as an information network of ideas and decisions.

2. Basic graph theory applied to web pages

A graph is a directed link. A link on a web page connects one document to another. A link is not only of a navigational nature but also can represent an endorsement to the target page. When we consider more than just one link, we could explore characteristics of the web space. Spatial relations between web pages can help understand the topology of a web page and in general of the web space. In the space of all web pages W, let $A \in W$ to mean a page A belongs to the space of all web pages. The web page A represents a graph. In that graph, if there is a link to another web page B, we can say that: A is related to B by the link. In symbolic terms we can write $(A, B) \in \Re$, where $\Re$ is the relation "point to". We can add the following observations on the relation $\Re$:

1. If every web page is related to itself, we say that $\Re$ is reflexive.
2. For all web pages X and Y that belong to W, if $(X, Y) \in \Re \Rightarrow (Y, X) \in \Re$. Web pages X and Y in that case represent mutual endorsement. Then the relation is then said to be symmetric.
3. For all web pages X and Y that belong to W, if $(X, Y) \in \Re \Rightarrow (Y, X) \notin \Re$. Web pages X and Y are linked in a unidirectional way. Then the relation is then said to be anti-symmetric.
4. For all web pages that belong to W, when a web page X cites another web page Y and that last page cites another web page Z, we can say that $\Re$ is transitive:

   $$(X, Y) \in \Re \quad \text{and} \quad (Y, Z) \in \Re \Rightarrow (X, Z) \in \Re.$$

5. When a web page cites X another web page Y and Y does not cite X, X endorses Y and Y does not endorse X, we can say that $\Re$ is not symmetric:

$$(X, Y) \in \Re \quad \text{but} \quad (Y, X) \notin \Re.$$

6. When two web pages X and Y point to a distinct third web page Z, then we could say that the two web pages are related through a very special relationship similar to a filtering relationship or bibliographic coupling (Kessler, 1963). This kind of relationship does not have a name in the algebraic properties of $\Re$

$$(X, Z) \in \Re \quad \text{and} \quad (Y, Z) \in \Re.$$

7. Conversely when one web page X points to two distinct web pages Y and Z, then we say that X co-cites Y and Z. Co-citation is a term borrowed from the field of bibliometric studies (Small, 1973). Co-citation has been used as a measure of similarity between web pages by Larson (1996) and Pitkow and Pirolli (1997). Small and Griffith (1997) used breadth-first search to compute the connected components of the uni-directed graphs in which two nodes are joined by an edge if and only if they have a positive co-citation value. This kind of relationship does not have a name in the algebraic properties of $\Re$

$$(X, Y) \in \Re \quad \text{and} \quad (X, Z) \in \Re.$$

These seven properties are the simplest common patterns that can be perceived on the web. These seven properties can blend together to form more complex patterns that are indicative of emerging links or communities on the web. These complex patterns can model properties of web pages that can be qualified as "authoritative web pages". An authoritative web page is a web page that is pointed at by many other web pages. Other emerging complex patterns can model web pages that can be qualified as survey web pages or "hub web pages" because they cite authoritative web pages.

2.1. Adjacency matrix representation of a web graph

To further apply all these properties, consider a directed graph G that represents a hyperlinks between web pages (Kleinberg et al., 1999). Consider also its adjacency matrix A. An entry a_{pq} is defined by the following:

$$a_{pq} = \begin{cases} 1 & \text{if there is an edge or link between two web pages } p \text{ and } q, \\ 0 & \text{otherwise.} \end{cases}$$

Here some of the properties that could be discovered from an adjacency matrix perspective:

1. A graph is said to be reflexive if every node in a graph is connected back to itself, i.e., $a_{pp} = 1$. The situation will happen if a page points back to itself.
2. A graph is said to be symmetric if for all edges p and q in X: iff $a_{pq} = 1$ then $a_{qp} = 1$. We say in this case that there is mutual endorsement.
3. A graph is said to be not symmetric if there exists two edges p and q in G such that iff $a_{pq} = 1$ then $a_{qp} = 0$. We say in this case that there is endorsement in one direction.
4. A graph is said to be transitive if for all edges p, q, and r:

$$\text{Iff } a_{pq} = 1 \quad \text{and} \quad a_{qr} = 1 \quad \text{then } a_{pr} = 1.$$

We say in this case that all links p endorse links r even though not directly.

5. A graph is said to be anti-symmetric iff for all edges p and q:
   Iff $a_{pq} = 1$ then $a_{qp} = 0$.
6. If two different web pages p and q point to another web page r then we say that there is social filtering. This means that these web pages are related through a meaningful relationship
   $a_{pr} = 1$ and $a_{qr} = 1$.
7. If a single page p points to two different web pages q and r then we say that there is co-citation
   $a_{pq} = 1$ and $a_{pr} = 1$.

Now consider two linear transformations defined on unit vectors a and h as follows:

$$a = A^{\mathrm{T}}h, \tag{1}$$

$$h = Aa \tag{2}$$

thus

$$a = AA^{\mathrm{T}}a, \tag{3}$$

$$h = A^{\mathrm{T}}Ah. \tag{4}$$

By examining closely the entries of these product matrices AA^{T} and $A^{\mathrm{T}}A$. These two matrices are symmetric with the following properties observed:

1. An entry (p,p) in AA^{T} means the number of web pages that come out of p. We call that number the out-degree or *od*.
2. An entry (p,p) in $A^{\mathrm{T}}A$ means the number of web pages that point towards p. We call that number in-degree or *id*.
3. An entry (p,q) in $A^{\mathrm{T}}A$ represents the number of web pages that are in common between p and q that point towards p and q.
4. An entry (p,q) in AA^{T} represents the number of web pages that came out of p and q that are in common.

   To illustrate the above points let us look at the following graph G.

   Here are the algebraic properties of $\Re$ in G:
   1. $\Re$ is not reflexive;
   2. $\Re$ is not symmetric;
   3. $\Re$ is not transitive;
   4. $\Re$ is anti-symmetric;
5. $(1,4) \in \Re$, $(3,4) \in \Re$, and $(5,4) \in \Re$: we could say that the vertex with the highest number of web pages pointing to it.
6. $(5,1) \in \Re$ and $(5,6) \in \Re$: 5 co-cites 1 and 6.

   Here is the corresponding adjacency matrix for Fig. 1.

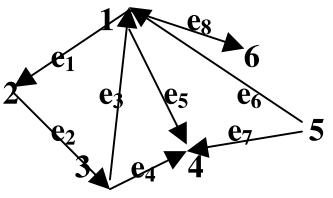

Fig. 1.

$$A = \begin{matrix} & & & & & & od \\ 0 & 1 & 0 & 1 & 0 & 1 & 3 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{matrix}$$

Building A^{T} yields

$$A^{\mathrm{T}} = \begin{matrix} 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{matrix}$$

Building $C = AA^{\mathrm{T}}$ yields

$$C = AA^{\mathrm{T}} = \begin{matrix} 3 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 2 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{matrix}$$

Building $D = A^{\mathrm{T}}A$ yields

$$D = A^{\mathrm{T}}A = \begin{matrix} 2 & 0 & 0 & 2 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 3 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \end{matrix}$$

Fig. 2 illustrates the in degrees and out degrees for the graph G:
How far away are two web pages in a web graph?

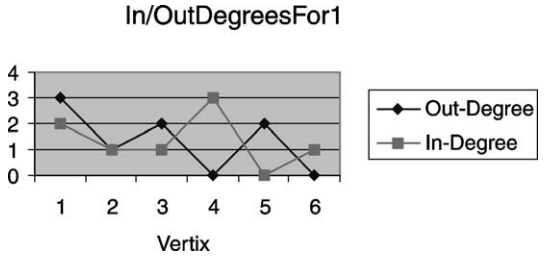

Fig. 2.

The adjacency matrix A can be used to calculate the length of the path than can separate two distinct web pages. To further explore such an idea, consider the power matrices of A, i.e., $A^2, A^3, A^4, \ldots, A^n$ for a graph of n vertices. If we calculate the value of A^2 for Graph1 we have:

$$A^2 = \begin{matrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{matrix}$$

Every non-zero element in A^2 means that to travel from vertex i to vertex j we will need two web links to get there. Thus considering that $A^2(2,4) = 1$ means that the distance from vertex 2 to vertex 4 is 2. This can be verified on Fig. 1 where $(2,3) \in \Re$ and $(3,4) \in \Re$.

If we calculate the rest of powers of A, i.e., $A^3, A^4, \ldots, A^n$, and we reach a value $m < n$ such that $A^m = A$ then we say that any two web pages in that graph are m pages or clicks away.

Applying this to Graph1, one can see that $A^4 = A$. This means that any two web pages the furthest away they are is four web pages. An example in Graph1 is web page 1 and 6, where to reach 6 from 1 we can travel directly with a path of length 1 or through vertices 2, 3, 1, 6.

Expanding the idea of the distances of web pages over the WWW, Albert, Jeong, and Barabasi (1999) were able to show that the distribution of web pages over the web constitutes a power law and that the distance between far away connected web pages is 19. In other words, to move along the whole WWW, it will take approximately 19 web pages or clicks at most. Thus the diameter of the WWW is 19 clicks.

2.2. Incidence matrix representation of a web graph

Another interesting representation of a graph is an incidence matrix. Let us consider the same graph G in Fig. 1 and its corresponding incidence matrix I. To build the incidence matrix of a graph we label the rows with the vertices and the columns with the edges (in any order). The entry i_{ve} for row r (vertex v) and column c (edge e) is such

$$i_{ve} = \begin{cases} 1 & \text{if } e \text{ is incident on } v, \\ 0 & \text{otherwise.} \end{cases}$$

Notice that id, which is the number of edges incident on a vertex v, can be deduced from the incidence matrix. We also added to I the row s which the sum of all the values in a given column

$$I = \begin{matrix} & e_1 & e_2 & e_3 & e_4 & e_5 & e_6 & e_7 & e_8 & id \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 2 \\ 2 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 3 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 4 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 3 \\ 5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 6 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ s & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & \end{matrix}$$

We could deduce from I that web page 4 or vertex 4 is the one with the highest incidence of links to it. Web page 4 is an authoritative web page since it is the web page with most links pointing to it. We can deduce through the value of s that there are no bi-directional links on this graph. That is why this graph is anti-symmetric.

One way to look at how matrix A and vector id relate is by considering the following matrix–vector multiplication $A^{\mathrm{T}}U$ where A^{T} is the transpose of A already computed in Section 2.1 and U is the Unit vector 1.

Applying $A^{\mathrm{T}}U$ to Graph1 results in the following:

$$A^{\mathrm{T}} \times U = \begin{vmatrix} 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{vmatrix} \times \begin{vmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{vmatrix} = \begin{vmatrix} 2 \\ 1 \\ 1 \\ 3 \\ 0 \\ 1 \end{vmatrix} \tag{5}$$

$$A^{\mathrm{T}} \times U = id.$$

The matrix I has been ignored in the literature on web graph theory (Kleinberg, 1998). Looking closely at II^{T} will yield the following matrix:

$$II^{\mathrm{T}} = \begin{matrix} 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{matrix}$$

II^{T} is a diagonal matrix. Its eigenvalues vector is equal to the in-degree vector id. Its eigenvectors constitute the columns of the unit matrix of size 6×6

$$\text{Eigenvalue } (II^{\mathrm{T}}) = id. \tag{6}$$

By looking at Eqs. (5) and (6) we can see how I and A are related

$$\text{Eigenvalue } (II^{\mathrm{T}}) = id = A^{\mathrm{T}} \times U. \tag{7}$$

2.3. Linear algebra and web graph theory

From linear algebra (Golub & Van Loan, 1989) if A is a symmetric $n \times n$ matrix, then an eigenvalue of A is a number g with the property for some vector v we have $Av = gv$. The set of all such v is a subspace of R^n, which we refer to as the eigenspace associated with g; the dimension of this space is the multiplicity of g. A has n distinct eigenvalues, each of them is a real number, and the sum of their multiplicity is n. These eigenvalues are denoted by $g_1(A), g_2(A), \ldots, g_n(A)$ indexed in order of decreasing absolute value and with each value listed in a number of times equal to its multiplicity. For each distinct eigenvalue, we choose an orthonormal basis of its eigenspace. Considering the vectors in all these bases, we obtain a set of eigenvectors $v_1(A), v_2(A), v_3(A), \ldots,$

$v_n(A)$ where we can say that $v_i(A)$ belongs to the eigenspace of $g_i(A)$. The following assumption can be made without any violation of any substantial principle in what follows that:

$$|g_1(A)| > |g_2(A)|. \tag{8}$$

When the last assumption holds Eq. (8), $v_1(A)$ is referred to as the principal eigenvector and all other $v_i(A)$ are non-principal eigenvectors.

By applying this analysis to II^{T} we can conclude:

1. There are four distinct eigenvalues mainly: 0, 1, 2, 3 with the multiplicity of $g = 1$ being 3.
2. The eigenspace can be defined by: $3(II)^{\mathrm{T}} > 2(II)^{\mathrm{T}} > 1(II)^{\mathrm{T}} \geqslant 1(II)^{\mathrm{T}} \geqslant 1(II)^{\mathrm{T}} > 0(II)^{\mathrm{T}}$.
3. There is one principal eigenvector which corresponds to the eigenvalue of $g = 3$. This vector has a value $v_3 = (0, 0, 0, 1, 0, 0)$. All other eigenvectors are not principal eigenvectors.

This analysis will be helpful in Section 4 on hubs and authorities.

2.4. Bipartite graphs

A bipartite graph G is a graph where the set of vertices can be divided into sets V_1 and V_2 such that each edge is incident on one vertex in V_1 and one vertex in V_2. Graph G in Fig. 1 is not an actual bipartite graph. To make G an actual bipartite graph, a possible bipartite graph G_1 can be designed.

If we let $V_1 = \{1, 3, 5\}$ and $V_2 = \{2, 4, 6\}$, then we can take out the two edges e_3 and e_7 that were in and then the new graph G_1 will become a bipartite graph (see Fig. 3).

In other related works, tree structures have been used to design better hyperlink structures (Botafogo, Rivlin, & Shneiderman, 1992). The reverse process of discovering tree structures from hyperlink web pages and discover hierarchical structures has also been studied (Mukherjea, Foley, & Hudson, 1995; Pirolli, Pitkow, & Rao, 1996).

In case of topic search, we do not need to extract a web structure from the web. Often the user is interested in finding a small number of authoritative pages on the search topic. These pages will play an important role in a tree had we extracted the tree structure itself. An alternative to extracting trees from a web graph is to use a ranking method to the nodes of the web graph. In this section we review such methods proposed in the literature. Some basic concepts have to be laid down before doing that.

We conclude that the topology of the links in web pages affect search performance and strategies of the WWW.

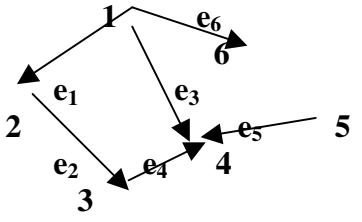

Fig. 3.

2.5. Web topology

In this section, we will explore how different can web graphs be? Can we classify these different web pages? How complex these web pages can appear to be?

Different categories of web graphs can emerge according to their *od* value and *id* value or what we is defined in Section 2.1 as out-degree and in-degree correspondingly. Even though we do not pretend that this classification is exhaustive, but different kinds of graphs were gathered to represent the possible different variety of web pages. Emerging web graphs can be complex and rich in structure and links more than web page designers do realize.

2.5.1. In-degree web graphs

Complex pages can emerge with large in-degree that look like Fig. 4. Fig. 5 illustrates such in-degree web pages by looking as their in/out degree chart.

2.5.2. Out-degree web graphs

Complex web pages with large out-degree can emerge that look like Fig. 6. Such graph becomes a tree where the starting node is the root of the tree:

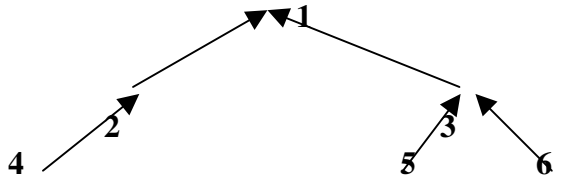

Fig. 4.

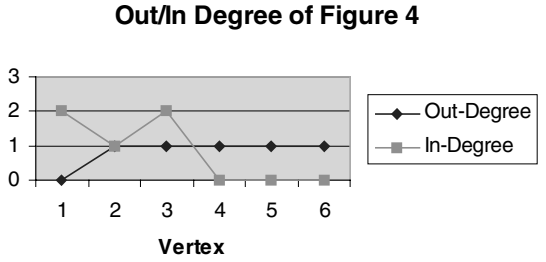

Fig. 5.

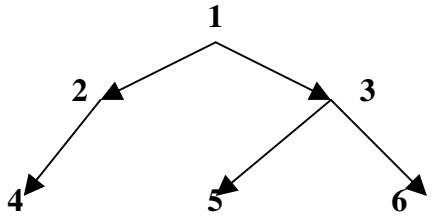

Fig. 6.

Fig. 7 illustrates such out-degree web pages by looking as their in/out degree chart, which is the complement of Fig. 5.

2.5.3. Complete bipartite web graphs

Other complex web pages can emerge as complete bipartite graphs that look like Fig. 8 with 3 nodes in the first set V_1 and 3 nodes in the second set V_2.

Remember that the topology of complete bipartite graphs like the one in Fig. 8 is unique (see Fig. 9).

2.5.4. Bipartite web graphs

Other complex web pages can emerge as bipartite graphs that look like Fig. 10 with 4 nodes in the first set V_1 and 2 nodes in the second set V_2.

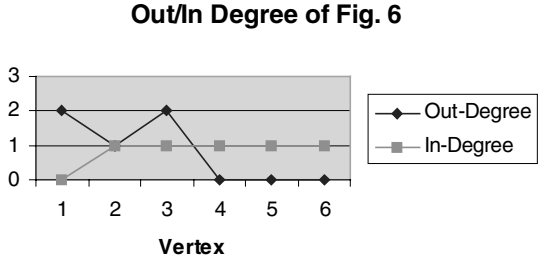

Fig. 7.

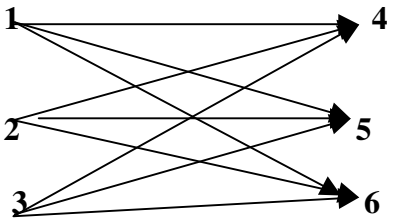

Fig. 8.

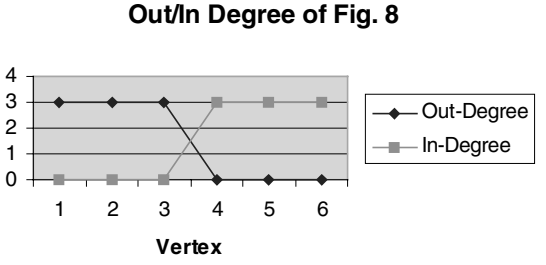

Fig. 9.

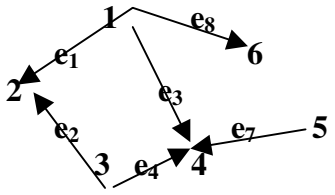

Fig. 10.

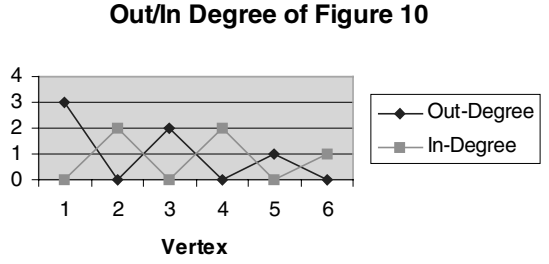

Fig. 11.

The difference between complete bipartite web graphs and bipartite graphs is the fact that not all nodes between set V_1 and V_2 are connected as it is seen in Fig. 10.

Pages with large in-degree or out-degree play an important role in web algorithms in general. Section 3 will apply Kleinberg's algorithm (Kleinberg, 1998) on these different graphs (see Fig. 11).

2.6. Web pages topological difference

The signature of a web graph lies in their in/out degree as it can be seen from all these different charts. The in/out degree of a graph can be used to measure the differences or similarities between different topological web structures. The signature of a web graph lies in that. The Euclidean distance will help classify different graphs.

Not only will be interested in the size of a graph but also the structure of the graph. A graph will be assumed to be symbolized by two vectors the in-degree *id* vector and the out-degree vector *od*. Next the average of the in-degree *id* vector will be calculated, the average value of the out-degree *od* vector will be calculated, and say that for a vertex v:

$$od(v) = \begin{cases} 1 & \text{if its value is above the average value,} \\ 0 & \text{if otherwise.} \end{cases} \tag{9}$$

$$id(v) = \begin{cases} 1 & \text{if its value is above the average value,} \\ 0 & \text{if otherwise.} \end{cases} \tag{10}$$

By applying (9) and (10) to Fig. 1, which is a general graph, we deduce

$$od = (1,0,1,0,1,0), \quad id = (1,0,0,1,0,0).$$

By applying (9) and (10) to Fig. 10, which is a possible bipartite graph on Fig. 1, we deduce

$$od = (1,0,1,0,1,0), \quad id = (0,1,0,1,0,1).$$

By applying (9) and (10) to Fig. 8, which is the only complete possible bipartite graph on Fig. 10, we deduce

$$od = (1,1,1,0,0,0), \quad id = (0,0,0,1,1,1).$$

By applying (9) and (10) to Fig. 4, which is an in-degree graph, we deduce

$$od = (0,1,1,1,1,1), \quad id = (1,1,1,0,0,0).$$

By applying (9) and (10) to Fig. 6, which is an out-degree graph, we deduce

$$od = (1,1,1,0,0,0), \quad id = (0,1,1,1,1,1).$$

The following matrix M summarizes the difference between these different web graphs with the same number of nodes in their (out-degree, in-degree) coordinate space

$$M = \begin{array}{c|ccccc} & \text{GG} & \text{BG} & \text{CBG} & \text{ID} & \text{OD} \\ \hline \text{GG} & (0,0) & (0,3) & (2,3) & (3,3) & (2,5) \\ \text{BG} & (0,3) & (0,0) & (2,2) & (3,4) & (2,2) \\ \text{CBG} & (2,3) & (2,2) & (0,0) & (4,6) & (0,2) \\ \text{ID} & (3,3) & (3,4) & (4,6) & (0,0) & (4,4) \\ \text{OD} & (2,5) & (2,2) & (0,2) & (4,4) & (0,0) \end{array}$$

The smallest value in this matrix is the value $(0,2)$, which says that Graphs 8 and 6 are the closest because a complete bipartite graph is a form of an out-degree graph with many roots. The next best smallest value in the matrix M is $(0,3)$ which says that general graphs and bipartite graphs are the closest among all other graphs. The largest value in M is $(4,6)$ which says that complete bipartite graphs and in-degree are the farthest apart. The next biggest difference is $(4,4)$ which says that the next largest difference is between in-degrees trees and out-degree trees, which is evident from the structure of the trees. It also shows that bipartite graphs are as close to out-degree trees and complete bipartite graphs than in-degree trees which can be concluded from the statement before.

We conclude that in the coordinate space of (out-degree, in-degree) the following metric of graphs topology stands:

$$|(\text{CBG})| < |(\text{OD})| < |(\text{BG})| < |(\text{GG})| < |(\text{ID})|, \tag{11}$$

where CBG = a complete bipartite graph, OD = out-degree trees, BG = bipartite graph, GG = general graphs, and ID = in-degree trees.

Fig. 12 displays the classification of the web graphs in the new coordinate (out-degree, in-degree) space. In Fig. 12, data1 represents all the data in the first row of matrix M, data 2 represents all the data in second row of matrix M, data3 represents all the data in the third row of matrix M, data4 represents all the data in the fourth row of matrix M, data5 represents all the data in the fifth row of matrix M.

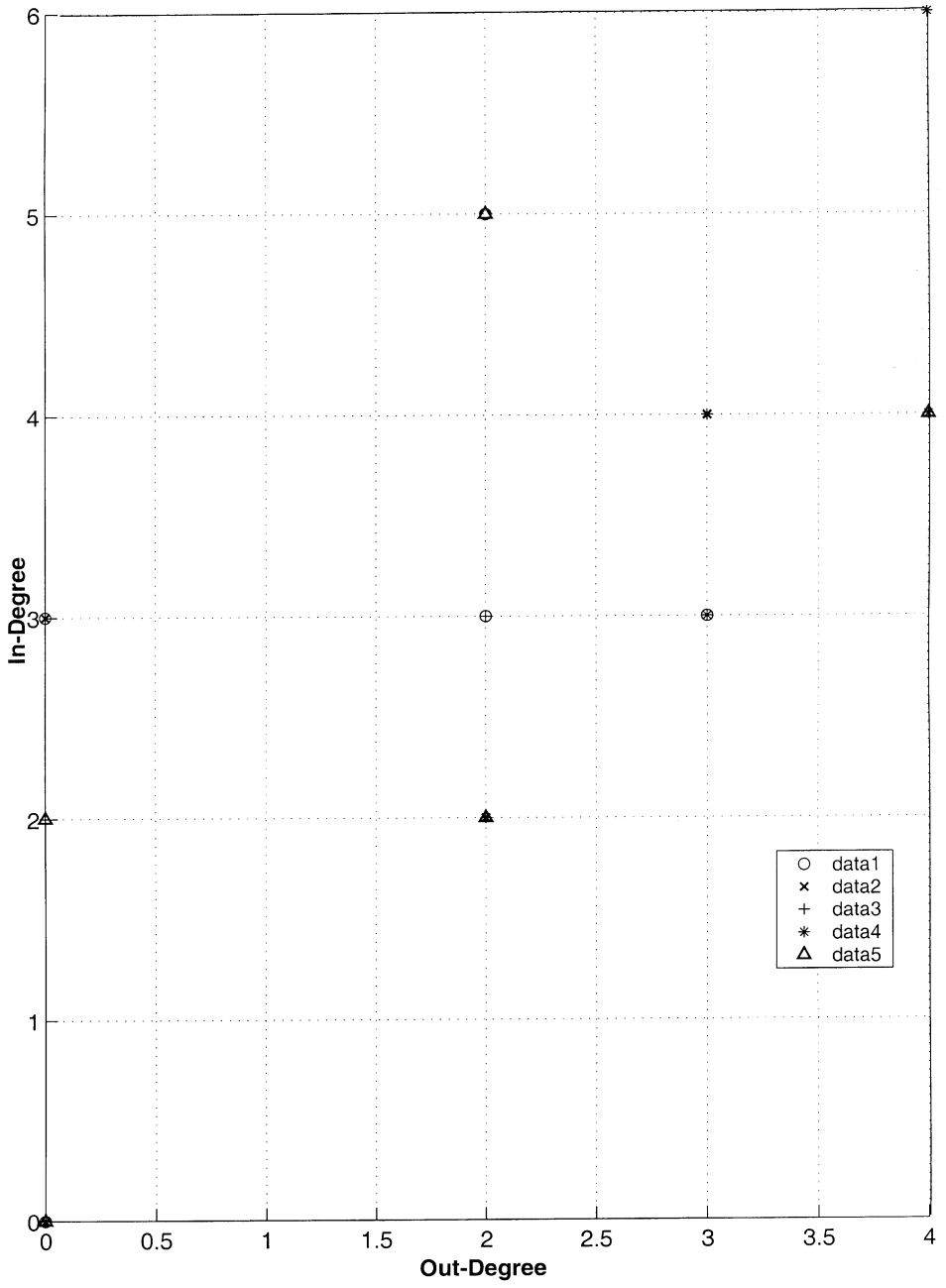

Fig. 12. Classification of web graphs in the (out-degree, in-degree) coordinate space.

3. Hubs and authorities

In 1998, Kleinberg built a search technique procedure which he called HITS, which stands for "Hyperlink Induced Topic Search". It was meant to rank the result set of links by any

commercial search engine on broad search topics. The first step in HITS is a sampling step. The result of such a sampling is called the root set. Each page tends to be considered a hub or an authority or neither. HITS try to find "hubs" web pages and "authorities" web pages.

An authority page is a page that has the highest number of links pointing to it in a given topology. A hub is a web page that points to a high number of authority web pages. Authorities and hubs reinforce each other mutually: since web pages tend to be part of what makes an authority or so close to an authority that they are hubs. There are pages that are neither authorities nor hubs. Good authority web pages are found close to good hubs, which in turn are pointing towards good sources of information.

A typical representation of Hubs and authorities can be seen as a bipartite graph. Fig. 15 represents an ideal situation of hubs and authorities. The root set mentioned above tends to contain few hundred web pages. To the root set is added links that point to it. This new set can grow to form up to few thousands of web pages and is called the base set. The second step in HITS is to compute the weight for each web page that can help rank the links as their relevance to the original query (see Fig. 13).

Even though HITS is applied to a large number of web pages not individual web pages, for illustration only, we can look at the web pages in Graph1. From the actual definition of authority web pages and hub web pages we can conclude:

1. Web page 4 is an "authority" web page because it has the highest number of pages that point to it. Web page 1 is also an authority web page.
2. Web pages 5 and 3 are good "hub" pages. Web pages 5 and 3 cannot be considered authorities.
3. Web pages 2 and 6 are neither hubs nor authorities.
4. Thus, a web page in HITS tends to have a hub weight and authority weight.

HITS assigns an authority weight a_p and a hub weight h_p for every web page that reinforce each other:

$$a_p = \sum_{(q,p)\in\Re} (h_q), \tag{12}$$

$$h_p = \sum_{(p,q)\in\Re} (a_q). \tag{13}$$

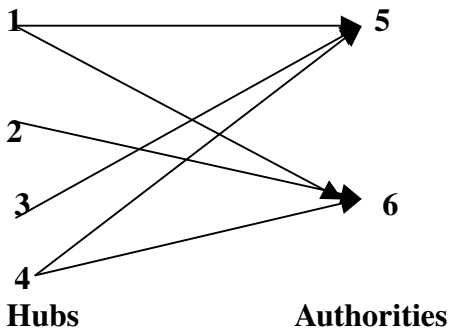

Fig. 13.

These values of a_p and h_p are maintained normalized on all the web pages for that particular graph. Thus

$$\sum_p (a_p)^2 = 1 \quad \text{and} \quad \sum_p (h_p)^2 = 1. \tag{14}$$

The larger the value of a_p the better representative of an authority is p. The higher the value of h_p the better representative of a hub is p. These values get initialized first to 1 and then get adjusted and the first few hubs and the first few authorities are then filtered out on a given graph. The procedure Filter looks like the following:

Filter F (Kleinberg, 1998):

1. Let z is a vector $z = (1, 1, 1, \ldots, 1)$ of R^n, where n is the number of web pages consider
2. Initialize a_{p0} and h_{p0} to z
3. For $i = 1, 2, \ldots, k$
   (a) Use (12) to (a_{pi-1}, h_{pi-1}) obtaining new a weights a'_i
   (b) Use (13) to (a'_{pi}, h_{pi-1}) obtaining new h weights h'_i
   (c) Normalize a'_{pi}, and get a_{pi}
   (d) Normalize h'_i, and get h_i
4. End
5. Return (a_{pk}, h_{pk})

Conjecture 1 (Kleinberg, 1998). *It turns out that the returned set (a_{pk}, h_{pk}) is made out of two values where a_{pk} is the principal eigenvector over $A^{\mathrm{T}}A$, and h_{pk} is the eigenvector over AA^{T}.*

Conjecture 2 (Kleinberg, 1998). *It turns out that that a web page cannot be an authority web page and a hub web page after the first iteration according to Kleinberg's Filter Procedure.*

Conjecture 3 (Kleinberg, 1998). *It turns out that that a web page cannot be an authority web page and a hub web page at the end of procedure Filter according to Kleinberg's Filter Procedure.*

3.1. HITS applied to general graphs

Let us apply the procedure Filter F to Graph1. The number of web pages in Graph1 is $n = 6$. Parameters a and h after one single iteration of step 3 in F and before any normalization have the following values: $a = (2, 1, 1, 3, 0, 1)$ and $h = (5, 1, 5, 0, 5, 0)$. After normalization $a = (0.5, 0.25, 0.25, 0.75, 0, 0.25)$ and $h = (0.5735, 0.1147, 0.5735, 0, 0.5735, 0)$. Here are some of the observations after the first iteration:

1. Web page 4 is the best authority web page.
2. Web pages 1, 3 and 5 are the best hub pages.
3. Web page 2 is the second best hub web page.
4. Web pages 2,3 and 6 are the third best authority web pages.
5. Web pages 4 and 6 are not hub web pages.
6. Web pages 1, 2, and 3 are authority and hub web pages at the same time with different ranks.

General graphs make appear the idea of a web page that is both an authority web page and a hub page in the first iteration. Web pages 1, 2, and 3 in Graph1 prove that to a certain extent.

Would that idea persist after a number of iterations? Conjecture 2 is not verified in the case of general graphs. What happens after a number of iterations? The same idea we used in the subsection on the "distance between web pages" can be applied here. Iterate until the values of vectors a and h do not change any more. What happened to Conjecture 3 in the case of general graphs?

Here are the final values of a and h after $k = 6$ iterations with a change of almost 0:

$a = (0.5164, 0.2582, 0, 0.7746, 0, 0.2582)$,

$h = (0.5774, 0, 0.5774, 0, 0.5774, 0)$.

Here are some final observations after $k = 6$ iterations in the procedure Filter:

1. Web page 4 is still the best authority web page.
2. Web pages 1, 3 and 5 are still the best and only hub web pages.
3. Web page 1 is now the second best authority web page.
4. Web pages 2 and 6 are now the third best authority web pages.
5. Web page 3 is not an authority web page any more.
6. Web page 1 is now the only hub web page and authority web page.

Web page 1 is still the key issue to tackle in the whole subject of hubs and authorities. It was not well expounded in the literature whether that is a major obstacle to the full expansion of hubs and authorities in general. Conjecture 3 does not hold for general graphs.

Being satisfied with the early set of iterations for a and h would have been deceiving. Some more values which had a non-zero value now have a zero value. The answer to that question lies in the fact that we want to verify that the final values of vectors a and h correspond to the eigenvectors of both $A^{\mathrm{T}}A$ and AA^{T} correspondingly.

From $D = A^{\mathrm{T}}A$ in Section 1 we can calculate the eigenvalues and eigenvectors of D. The columns of the following matrix M_1 constitute the eigenvectors:

0.0000	−0.1647	−0.6070	−0.0665	0.5774	0.5164
0.0000	−0.2578	−0.1820	−0.7074	−0.5774	0.2582
−1.0000	0.0000	−0.0000	−0.0000	−0.0000	−0.0000
−0.0000	0.1647	0.6070	0.0665	−0.0000	0.7746
−0.0000	0.9331	−0.2221	−0.2829	0.0000	0.0000
−0.0000	0.0931	−0.4250	0.6409	−0.5774	0.2582

Here is the corresponding eigenvalue vector:

1.0000
0.0000
0.0000
0.0000
2.0000
5.0000

So according to our notations in Section 2.3 we could say that

$$5(A^{\mathrm{T}}A) > 2(A^{\mathrm{T}}A) > 1(A^{\mathrm{T}}A) > 0(A^{\mathrm{T}}A) \geqslant 0(A^{\mathrm{T}}A) \geqslant 0(A^{\mathrm{T}}A).$$

So the eigenvalue is 5.000 which is the sixth value in the vector of eigenvalue. The corresponding eigenvector is a vector denoted by $w_6(M_1)$ which the last column of M_1 would be:

0.5164
0.2582
0.0000
0.7746
0.0000
0.2582

By comparing the last vector and the vector a already calculated above we can conclude that

$$w_6(M_1) = a. \tag{15}$$

From $C = AA^{\mathrm{T}}$ in Section 1 we can calculate the eigenvalues and eigenvectors of C. The columns of the following matrix M_2 constitute the eigenvectors:

0	0	0	0.8165	0.5774	0
1	0	0	0	0	0
0	0.2571	−0.6587	−0.4082	0.5774	0
0	0.9316	0.3635	−0.0000	0.0000	0
0	−0.2571	0.6587	−0.4082	0.5774	0
0	0	0	0	0	1

Here is the corresponding eigenvalue vector

1.0000
0.0000
0.0000
2.0000
5.0000
0.0000

So according to our notations in Section 2.3 we could say that

$$5(AA^{\mathrm{T}}) > 2(AA^{\mathrm{T}}) > 1(AA^{\mathrm{T}}) > 0(AA^{\mathrm{T}}) \geqslant 0(AA^{\mathrm{T}}) \geqslant 0(AA^{\mathrm{T}}).$$

So the eigenvalue is 5.000 which is the fifth value in the vector of eigenvalue. So the principal eigenvector denoted by $w_5(M_2)$ which the fifth column of D would be

0.5774
0
0.5774
0
0.5774
0

By comparing the last vector and the vector h already calculated above we can conclude that

$$w_5(M_2) = h. \tag{16}$$

3.2. HITS applied to in-degree graphs

Let us apply the Filter procedure to an in-degree graph like the one in Fig. 14. The number of web pages in Fig. 14 is $n = 8$. Parameters a and h after one single iteration of step 3 in F and before any normalization have the following values: $a = (0, 0, 0, 0, 0, 2, 2, 3)$ and $h = (2, 2, 3, 2, 2, 3, 3, 0)$. After normalization $a = (0, 0, 0, 0, 0.4851, 0.4851, 0.7276)$ and $h = (0.3050, 0.3050, 0.4575, 0.3050, 0.3050, 0.4575, 0.4575, 0)$. Here some of the observations that can be made from a and h after one single iteration ($k = 1$):

1. Web page 8 is the best authority web page.
2. Web pages 6 and 7 are the second best authority web pages.
3. Web pages 3, 6 and 7 are the best hub web pages.
4. Web pages 1, 2, 4, and 5 are the second best hub web pages.
5. Web pages 6 and 7 are authority web pages and hub web pages at the same time.

The first iteration of the procedure Filter shows a web page that is both an authority web page and a hub page. Web pages 6 and 7 in that example show clearly that. Conjecture 2 does not hold for in-degree graphs. Would that idea persist after a number of iterations? What happened to Conjecture 3 in the case of in-degree graphs.

Here are the final values of a and h after $k = 8$ iterations with a change of almost 0:

$a = (0, 0, 0, 0, 0, 0.039, 0.039, 0.9985)$,

$h = (0.02, 0.02, 0.5768, 0.02, 0.02, 0.5768, 0.5768, 0)$.

Here are the final observations after $k = 8$ iterations on the procedure Filter:

1. Web page 8 is the only authority web page.
2. Web pages 3, 6, and 7 are the only hub pages.

The process of identifying hubs and authorities is an iterative process and the first iteration is just the beginning of the process of filtering out weak hubs and authorities. In-degree trees seem to be ideal for the procedure Filter since no one web page can be considered both a hub and an authority at the same time. Conjecture 3 holds for in-degree graphs.

Being satisfied with the early set of iterations for a and h would have been deceiving. Some more values which had a non-zero value now have a zero value. The answer to that question lies in the fact that we want to verify that the final values of vectors a and h correspond to the eigenvectors of both $A^{\mathrm{T}}A$ and AA^{T} correspondingly.

We want still to verify Conjecture 1. It turns out that $D = A^{\mathrm{T}}A$ is a diagonal matrix. The eigenvalues and eigenvectors can be calculated simply with no other transformations like what we did in the case of general graphs

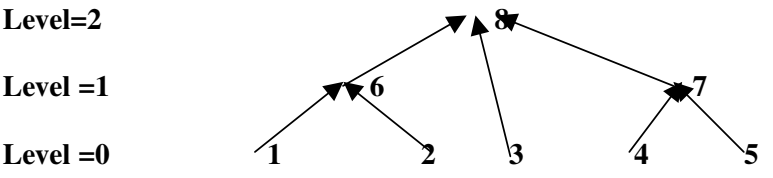

Fig. 14.

$$\begin{matrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 3 \end{matrix}$$

The corresponding matrix M_1 contains the eigenvectors is the unit matrix of 8×8. Here is the corresponding eigenvalue vector:

$$\begin{matrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 2 \\ 2 \\ 3 \end{matrix}$$

So according to our notations in Section 2.3 we could say that:

$$3(A^{\mathrm{T}}A) > 2(A^{\mathrm{T}}A) > 2(A^{\mathrm{T}}A) > 0(A^{\mathrm{T}}A) \geqslant 0(A^{\mathrm{T}}A) \geqslant 0(A^{\mathrm{T}}A).$$

So the eigenvalue is 5.000 which is the eighth value in the vector of eigenvalue. So the principal eigenvector denoted by $w_8(M_1)$ which the last column of M_1 would be

$$\begin{matrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{matrix}$$

By comparing the last vector and the vector a already calculated above we can conclude that

$$w_8(M_1) \sim a. \tag{17}$$

From $C = AA^{\mathrm{T}}$, we can calculate the eigenvalues and eigenvectors of C. The columns of the following matrix M_2 constitute the eigenvectors:

0.7071	0.7071	0	0	0	0	0	0
−0.7071	0.7071	0	0	0	0	0	0
0	0	0	0	0.7850	0.2246	0.5774	0
0	0	−0.7071	0.7071	0	0	0	0
0	0	0.7071	0.7071	0	0	0	0
0	0	0	0	−0.5870	0.5676	0.5774	0
0	0	0	0	−0.1980	−0.7921	0.5774	0
0	0	0	0	0	0	0	1

Here is the corresponding eigenvalue vector

0
2
0
2
0
0
3
0

So according to our notations in Section 2.3 we could say that

$$3(AA^{\mathrm{T}}) > 2(AA^{\mathrm{T}}) > 2(AA^{\mathrm{T}}) > 0(AA^{\mathrm{T}}) \geqslant 0(AA^{\mathrm{T}}) \geqslant 0(AA^{\mathrm{T}}) \geqslant 0(AA^{\mathrm{T}}).$$

So the eigenvalue is 3 which is the seventh value in the vector of eigenvalue. So the principal eigenvector denoted by $w_7(M_2)$ would be

0
0
0.5774
0
0
0.5774
0.5774
0

By comparing the last vector and the vector h already calculated above we can conclude that:

$$w_7(M_2) \sim h. \tag{18}$$

3.3. HITS applied to out-degree graphs

Let us apply the Filter procedure to an out-degree graph like the one in Fig. 15. The number of web pages in Fig. 15 is $n = 8$. Parameters a and h after one single iteration of step 3 in F and before any normalization have the following values: $a = (0, 1, 1, 1, 1, 1, 1, 1)$ and $h = (2, 3, 2, 0,$

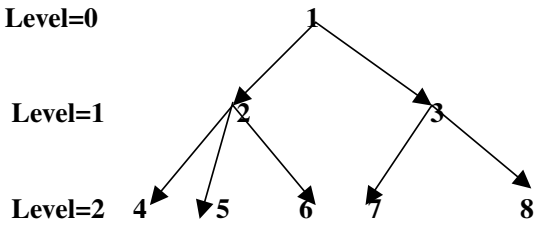

Fig. 15.

$0,0,0,0)$. The values of a and h after normalization are: $a = (0, 0.378, 0.378, 0.378, 0.378, 0.378, 0, 0.378, 0.378)$ and $h = (0.4581, 0.7276, 0.4581, 0, 0, 0, 0, 0)$.

Here are some observations that can be made from the first iterative values of a and h:

1. Web pages 2, 3, 4, 5, 6, 7, and 8 are authority pages.
2. Web pages 2 is the best hub web page.
3. Web pages 1 and 3 are the second best hub web pages.
4. Web pages 2 and 3 are hub web pages and authority web pages at the same time.

The first iteration of the procedure Filter shows a web page that is both an authority web page and a hub page. Web pages 2 and 3 in that example show clearly that. Conjecture 2 does not hold for out-degree graphs. Would that idea persist after a number of iterations? What happened to Conjecture 3 in the case of out-degree graphs?

Here are the final values of a and h after $k = 8$ iterations with a change of almost 0:

$a = (0.03, 0.03, 0.03, 0.576, 0.576, 0.576, 0.03, 0.03)$,

$h = (0.04, 0.999, 0.04, 0, 0, 0, 0, 0)$.

Here are some final observations after $k = 8$ iterations on the procedure Filter:

1. Web page 4, 5, and 6 are the only authority web pages.
2. Web page 2 is the only hub page.

This final result requires some interpretations. Even though web page 3 could have been considered a hub web page but because in the same graph more nodes came out of web page 2 than web page 3, we could say that the measure of a hub for a web page is a global measure and not a local one. Second, that because 2 is the nub page, web pages 4, 5, and 6 become the only authority web pages and not any more 7 and 8 like it was in the first iteration of the procedure Filter. Again the idea of an authority web page is a global measure and not a local one. This idea could be applied to citation analysis in general, which says that in a given literature on a given topic, being an authority is not a second best but really the best. The more you are cited, the more links come to your web page and the more attention you will receive, and that make you an authority in that field regardless of the second best. Further studies have reflected on the fact that there may be more than just authority in a given graph, which should be considered.

Conjecture 3 does hold for out-degree graphs.

Being satisfied with the early set of iterations for a and h would have been deceiving. Some more values which had a non-zero value now have a zero value. The answer to that question lies in the fact that we want to verify that the final values of vectors a and h correspond to the eigenvectors of both $A^{\mathrm{T}}A$ and AA^{T} correspondingly.

To verify Conjecture 1, we calculate $D = A^{T}A$. We can calculate the eigenvalues and eigenvectors of D. The columns of the following matrix M_1 constitute the eigenvectors:

1.0000	0	0	0	0	0	0	0
0	0.7071	0.7071	0	0	0	0	0
0	0.7071	−0.7071	0	0	0	0	0
0	0	0	0.7850	0.2246	0.5774	0	0
0	0	0	−0.5870	0.5676	0.5774	0	0
0	0	0	−0.1980	−0.7921	0.5774	0	0
0	0	0	0	0	0	0.7071	0.7071
0	0	0	0	0	0	0.7071	−0.7071

Here is the corresponding eigenvalue vector

0
2
0
0
0
3
2
0

So according to our notations in Section 2.3 we could say that

$$3(AA^{T}) > 2(AA^{T}) > 2(AA^{T}) > 0(AA^{T}) \geqslant 0(AA^{T}) \geqslant 0(AA^{T}) \geqslant 0(AA^{T}) \geqslant 0(AA^{T}).$$

So the eigenvalue is 3 which is the sixth value in the vector of eigenvalue. So the principal eigenvector denoted by $w_6(M_1)$ would be the principal eigenvector. So the principal eigenvector denoted by $w_6(M_1)$

0
0
0
0.5774
0.5774
0.5774
0
0

By comparing the last vector and the vector a already calculated above we can conclude that

$$w_6(M_1) \sim a. \tag{19}$$

It turns out that $C = AA^{T}$ is a diagonal matrix. The eigenvalues and eigenvectors can be calculated simply with no other transformations like what we did in the case of general graphs.

$$\begin{matrix} 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{matrix}$$

The corresponding matrix M_2 contains the eigenvectors is the unit matrix of 8×8. Here is the corresponding eigenvalue vector:

$$\begin{matrix} 2 \\ 3 \\ 2 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{matrix}$$

So according to our notations in Section 2.3 we could say that

$$3(A^{\mathrm{T}}A) > 2(A^{\mathrm{T}}A) > 2(A^{\mathrm{T}}A) > 0(A^{\mathrm{T}}A) \geqslant 0(A^{\mathrm{T}}A) \geqslant 0(A^{\mathrm{T}}A) \geqslant 0(A^{\mathrm{T}}A).$$

So the eigenvalue is 3 which is the second value in the vector of eigenvalue. So the principal eigenvector denoted by $w_2(M_2)$ would be the principal eigenvector. So the principal eigenvector denoted by $w_2(M_2)$

$$\begin{matrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{matrix}$$

By comparing the last vector and the vector h already calculated above we can conclude that

$$w_2(M_2) \sim h. \tag{20}$$

3.4. HITS applied to complete bipartite graphs

Let us apply the Filter procedure to a "complete bipartite" Graph like the one in Fig. 16. The number of web pages in Fig. 16 is $n = 7$. Parameters a and h after one single iteration of step 3 in F and before any normalization have the following values: $a = (0, 0, 0, 0, 4, 4, 4)$ and $h = (12, 12, 12, 12, 0, 0, 0)$. After normalization $a = (0, 0, 0, 0, 0.5774, 0.5774, 0.5774)$ and $h = (0.5, 0.5, 0.5, 0.5, 0, 0, 0)$. The following observations can be made from a and h after the first iteration:

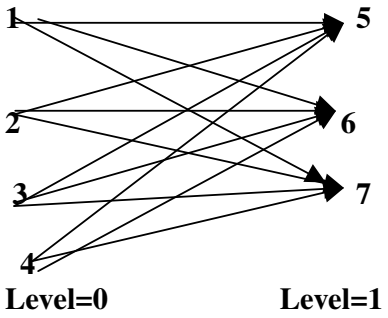

Fig. 16.

1. Web pages 5, 6, and 7 are the only authority pages.
2. Web pages 1, 2, 3 and 4 are the only hub pages.

In this perfect case no pages are either authority and hub pages at the same time. In this case also, there are no web pages that are neither authority web pages nor hub web pages. The observations made in the first iteration are evident from the complete bipartite graph itself. Conjecture 2 holds for complete bipartite graphs. What happened to Conjecture 3 in the case of complete bipartite graphs?

Complete bipartite graphs are ideal cases for finding hubs and authorities. We need not iterate any further the procedure Filter since no change will occur any way. But still one can iterate to verify that the values of a and h will stay the same. Here are the final values of a and h after $k = 8$ iterations with no change: $a = (0, 0, 0, 0, 0.5774, 0.5774, 0.5774)$ and $h = (0.5, 0.5, 0.5, 0.5, 0, 0, 0)$.

The early observations made on a and h will stand after a number of iterations.

Conjecture 3 holds for bipartite graphs.

3.5. HITS applied to bipartite graphs

Let us apply the Filter procedure to a bipartite graph like the one in Fig. 3. The number of web pages in Fig. 3 is $n = 6$. Parameters a and h after one single iteration of step 3 in F and before any normalization have the following values: $a = (0, 2, 0, 2, 0, 1)$ and $h = (3, 0, 4, 0, 2, 0)$. After normalization the values of a and h are: $a = (0, 0.6667, 0, 0.6667, 0, 0.3333)$ and $h = (0.5571, 0, 0.7428, 0, 0.3714, 0)$. The following observations can be made from the first values of a and h:

1. Web pages 2 and 4 are the best authority pages.
2. Web page 3 is the best hub page.
3. Web page 6 is the second best authority web page.
4. Web page 1 is the second best hub web page.
5. Web page 5 is the third best hub web page.
6. No web pages are authority web pages and hub pages at the same time.
7. All web pages are either authority web pages or hub web pages.

Conjecture 2 holds for bipartite graphs. What happened to Conjecture 3 for bipartite graphs?

Here are the final values of a and h after $k = 12$ iterations with almost no change:

$a = (0, 0.7370, 0, 0.5910, 0, 0.3280)$,

$h = (0.5910, 0, 0.7374, 0, 0.3280, 0)$.

The final observations can be made from the final values of a and h:

1. Web page 2 is the only best authority web page.
2. Web page 3 is still the best hub web page.
3. Web page 4 becomes the second best authority web page
4. Web page 6 is the third best authority web page.
5. Web page 1 is still the second best hub web page.
6. Web page 5 is the still third best hub web page.
7. No web pages are authority web pages and hub pages at the same time.
8. All web pages are either authority web pages or hub web pages.

Conjecture 3 holds for bipartite graphs.

Being satisfied with the early set of iterations for a and h would have been deceiving. Some more values which had a non-zero value now have a zero value. The answer to that question lies in the fact that we want to verify that the final values of vectors a and h correspond to the eigenvectors of both $A^{\mathrm{T}}A$ and AA^{T} correspondingly.

To verify Conjecture 1, we calculate $D = A^{\mathrm{T}}A$. Also we need to calculate the eigenvalues and eigenvectors of D. The columns of the following matrix M_1 constitute the eigenvectors:

1.0000	0	0	0	0	0
0	0	0	−0.7370	−0.3280	0.5910
0	0	1.0000	0	0	0
0	0	0	−0.5910	0.7370	−0.3280
0	1.0000	0	0	0	0
0	0	0	−0.3280	−0.5910	−0.7370

Here is the corresponding eigenvalue vector

0
0
0
3.247
1.555
0.1981

So according to our notations in Section 2.3 we could say that

$$3.247(AA^{\mathrm{T}}) > 1.555(AA^{\mathrm{T}}) > 0.1981(AA^{\mathrm{T}}) > 0(AA^{\mathrm{T}}) \geqslant 0(AA^{\mathrm{T}}) \geqslant 0(AA^{\mathrm{T}}).$$

So 3.247 is the eigenvalue which is the fourth value in the vector of eigenvalue. So the principal eigenvector would be $w_4(M_1)$. The value of the principal eigenvector denoted by $w_4(M_1)$ would be

0
−0.737
0
−0.5910
0
−0.3280

The reflections made before on using just the value v_i which corresponds to g_i is contested here. We should use the absolute value of $v_i(w)$ as we did with $g_i(A)$. By comparing the last vector and the vector a already calculated above we can conclude that

$$|w_4(M_1)| = a. \tag{21}$$

From $C = AA^{\mathrm{T}}$, we can calculate the eigenvalues and eigenvectors of C. The columns of the following matrix M_2 constitute the eigenvectors:

0	0	−0.7370	0.3280	0.5910	0
0	1.0000	0	0	0	0
0	0	0.3280	−0.5910	0.7370	0
1.0000	0	0	0	0	0
0	0	0.5910	0.7370	0.3280	0
0	0	0	0	0	1.0000

Here is the corresponding eigenvalue vector

0
0
1.555
0.1981
3.247
0

So according to our notations in Section 2.3 we could say that

$$3.247(AA^{\mathrm{T}}) > 1.555(AA^{\mathrm{T}}) > 0.1981(AA^{\mathrm{T}}) > 0(AA^{\mathrm{T}}) \geqslant 0(AA^{\mathrm{T}}) \geqslant 0(AA^{\mathrm{T}}).$$

So the eigenvalue is 3.247 which is the fifth value in the vector of eigenvalue. So the principal eigenvector would be $w_5(M_2)$. The value of the principal eigenvector denoted by $w_5(M_2)$ would be

0.591
0
0.737
0
0.328
0

By comparing the last vector and the vector h already calculated above we can conclude that

$$w_5(M_2) \sim h. \tag{22}$$

Note: All the above mathematic calculations were done in Matlab.

4. Final observations on hubs and authorities and conclusion

Conjecture 2 was true for some graphs and not true for other graphs. Web pages that started out to be both hub web pages and authority web pages in the early stages of the procedure Filter were soon changed into either hub web pages or authority web pages but not both any more at the

end of iterations. Conjecture 3 held ground for all graphs with the exception of general web graphs. The idea that a web page can be both a hub page and an authority web page is intriguing and worth further consideration. According to Kleinberg (2000) "nothing in the algorithm prevents this from happening but it is not a very common phenomenon".

Does our classification of these different web graphs developed in Eq. (11) in Section 2.6 help shed a light on the behavior of Kleinberg's algorithm (Kleinberg, 1998) even though Kleinberg's algorithm is not related to the (out-degree, in-degree) space in which all these different graphs have been studied. According to Eq. (11) complete bipartite graphs (CBG) have a behavior close to out-degree graphs (OD), and that OD graphs have a behavior close to general graphs (GG), and that GG have a behavior close to in-degree graphs (ID). Conjecture 2 held for CBG, BG but not OD, GG, and ID. Conjecture 2 violated the classification scheme established in that space in the case of OD but held it in the case of GG and ID. Conjecture 3 held for CBG, OD, BG, ID but not GG. According to (11) Conjecture 3 should not have held ground for ID given the fact that their behavior is close to GG than the other graphs. The general classification scheme proposed was able to help predict to a certain extent the behavior of these graphs. New classification schemes are needed to help predict the behavior of other web algorithms in the new space. The author is working on developing new classification schemes with other distances in the same (out-degree, in-degree) space.

Furthermore, in this study only the influence of principal eigenvectors on the selection of hubs and authorities was considered. What happened when we expand the role of these secondary eigenvectors? We will have not only have just a set of hubs and authorities but a multiple sets of hubs and authorities. How would the topological structure of the different kind of graphs studied in this paper influence the idea of a multiple sets of hubs and authorities? The author is in the process to expanding the role of these secondary eigenvectors in different topological structures.

Our study focused on just using links as a mean to evaluate web pages and uncover hubs and authorities. No heuristics or any other contextual information was used to further enhance the idea of hubs and authorities. In an early study, McBryan (1994) used searching hyperlinks based on an anchor text, in which one treats the text surrounding a hyperlink as a descriptor of the page being pointed to when assessing the relevance of that web page. Frisse (1997) considered the problem of document retrieval in single-authored, stand-alone works of hypertext. He proposed heuristics by which hyperlinks can enhance notions of relevance (Van Rijsbergen, 1979), and hence the performance of retrieval heuristics. In recent studies (Bharat & Henzinger, 1998; Chakrabarti et al., 1998a,b), three distinct user studies were performed to help evaluate the HITS system to evaluate information found on the WWW. Each one of the studies employed additional heuristics to further enhance relevance judgments. As such, these three studies cannot enhance the direct evaluation of the pure link-based method described here; rather they assess its performance as the core component of a WWW search tool. For example, in (Chakrabarti et al., 1998a), the CLEVER system was used to create an automatic resource compilation or the construction of lists of high-quality WWW pages related to a broad search topic and the goal was to see whether the output of CLEVER compared to that of a manually generated compilation such as the WWW search service of Yahoo for a set of 26 topics. A collection of 37 users was assembled; the users were required to be familiar with the use of a web browser, but were not experts in the topics picked. The users were asked to judge each web page as "bad", "fair", "good" or "fantastic" in terms of their utility of learning about the topic. For approximately 31% of the topics, the

evaluations of Yahoo and CLEVER were equivalent to within a threshold of statistical significance; for approximately 50% of the topics CLEVER was evaluated higher; and for the remaining 19% Yahoo was evaluated higher. Many of the users of these studies reported that they used the lists as starting points from which to explore, but that they visited many pages not on the original topic lists generated by the various techniques.

Of course it is hard to draw definitive conclusions from these three studies. We believe more insight in the structure of the graphs under consideration will help improve the success of web algorithms, i.e., HITS and CLEVER, to refine the original concepts and make place for better understanding of hubs and authorities in a variety of topological structures. This study does not pretend to have improved finding information on the WWW. This study focuses on discovering important topological structures in web pages and to predict the behavior of web algorithms in such environment especially that the WWW is rich not only in information but in topological structures.

In a continuing effort to further study ranking in different topological structures, the author has applied Google's ranking algorithm (Brin & Page, 1998; Brin, 1998) to these different structures (Meghabghab, 2001b). Early results show how Google's algorithm failed in different topological structures and had to be adapted to really fit the structure under consideration.

Last of all, what if a user's web page did not fit in any of the different web graphs already described? Can we evaluate such an activity in a given web page? May be different types of graphs are needed for that?

Our study raised a number of questions:

1. How do we categorize existing simple graphs such the one already in use in many areas of research?
2. How do we uncover web algorithms that are efficient on such graphs?
3. How do we devise new graphs, to better characterize user creativity in a given web page on the WWW?
4. How do we devise new algorithms for these graphs?

The WWW is full of information and mining the web for different kinds of information is still in its infancy. The difficulty in evaluating information on the web has to do with our cognitive abilities combined with personal likings, which is hard to quantify always. We are on the verge of unveiling creativity as a graph and what best describes the birth and death of an idea which happens in the birth and death of web pages and the changes of their content.

References

Albert, R., Jeong, H., & Barabasi, A. L. (1999). Diameter of the world wide web. *Nature*, *401*(9), 130–131.

Bharat, K., & Henzinger, M. R. (1998). Improved algorithms for topic distillation in a hyper-linked environment. In *Proceedings of the ACM conference on research and development in information retrieval*.

Brin, S. (1998). Extracting patterns and relations from the World Wide Web. In *Proceedings of WebDB'98*, Valencia, Spain.

Brin, S., & Page, L. (1998). The anatomy of a large scale hyper-textual web search engine. In *Proceedings of the seventh World Wide Web conference*, Brisbane, Australia.

Botafogo, R., Rivlin, E., & Shneiderman, B. (1992). Structural analysis of hypertexts: identifying hierarchies and useful metrics. *ACM Transactions on Information Systems*, *10*, 142–180.

Chakrabarti, S., Dom, B., Gibson, D., Kleinberg, J., Raghavan, P., & Rajagopalan, S. (1998a). Automatic resource compilation by analyzing hyperlink structure and associated text. In *Proceedings of the seventh international World Wide Web conference*, Brisbane, Australia.

Chakrabarti, S., Dom, B., Gibson, D., Kumar, S. R., Raghavan, P., Rajagopalan, S., & Tomkins, A. (1998b). Experiments in topic distillation. In *Proceedings of the ACM SIGIR workshop on hypertext information retrieval on the Web*.

Egghe, L., & Rousseau, R. (1990). *Introduction to informetrics*. Amsterdam: Elsevier.

Frisse, M. E. (1997). Searching for information in a hypertext medical handbook. *Communications of the ACM*, *31*(7), 880–886.

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

Golub, G., & Van Loan, C. F. (1989). *Van matrix computations*. Baltimore, MD: Johns Hopkins University Press.

Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American Documentation*, *14*, 10–25.

Kleinberg, J. (1998). Authoritative sources in a hyper-linked environment. In *Proceedings of the ACM-SIAM symposium on discrete algorithms* (pp. 668–677).

Kleinberg, J., Kumar, R., Raghavan, P., Rajagopalan, P., & Tomkins, A. (1999). The Web as a graph: measurements, models and methods. In *Invited survey at the international conference on combinatorics and computing*.

Kleinberg, J. (2000). Personal e-mail to the author.

Kopetzky, T., & Kepler, J. (1999). Visual preview for link traversal on the WWW. In *Proceedings of the eighth international World Wide Web'99* (*WWW8*) *conference*, Toronto, Canada.

Larson, R. (1996). Bibliometrics of the World Wide Web: an exploratory analysis of the intellectual structure of cyberspace, In *Annual meeting of the American Society Information Sciences*.

McBryan, O. (1994). GENVL and WWW: tools for taming the Web. In *Proceedings of the first international World Wide Web conference*.

Meghabghab, G. (2000). Stochastic simulations of rejected World Wide Web pages. In *Proceedings of the eighth IEEE international symposium on modeling, analysis and simulation of computer and telecommunications systems*, (*MASCOTS 2000*), August 29–September 1 (pp. 483–491), San Francisco, CA.

Meghabghab, G. (2001a). Iterative radial basis functions neural networks as metamodels of stochastic simulations of the quality of search engines in the World Wide Web. *Information Processing and Management*, 37, 571–591.

Meghabghab, G. (2001b). Google's web page ranking applied to different topological Web graph structures. *Journal of American Society for Information Sciences* (in press).

Mukherjea, S., Foley, J., & Hudson, S. (1995). Visualizing complex hypermedia networks through multiple hierarchical views. In *Proceedings of ACM SIGCHI conference in human factors in computing* (pp. 331–337), Denver, CO.

Pinski, G., & Narin, F. (1976). Citation influence for journal aggregates of scientific publications: theory with application to the literature of physics. *Information Processing and Management*, *12*, 297–312.

Pirolli, P., Pitkow, J., & Rao, R. (1996). Silk from a Sow's Ear: extracting usable structures from the web. In *Proceedings of ACM SIGCHI conference in human factors in computing*.

Pitkow, J., & Pirolli, P. (1997). Life, death, and lawfulness on the electronic frontier. In *Proceedings of ACM SIGCHI conference on human factors in computing*.

Small, H. (1973). Co-citation in the scientific literature: a new measure of the relationship between two documents. *Journal of the American Society for Information Sciences*, *24*, 265–269.

Small, H. (1986). The synthesis of specialty narratives from co-citation clusters. *Journal of the American Society for Information Sciences*, *37*, 97–110.

Small, H., & Griffith, B. C. (1997). The structure of the scientific literatures I. Identifying and graphing specialties. *Science Studies*, *4*, 17–40.

Van Rijsbergen, C. J. (1979). *Information retrieval*. London: Butterworths.