Google's Web Page Ranking Applied to Different Topological Web Graph Structures

George Meghabghab
Roane State, Department of Computer Science Technology, Oak Ridge, TN 37830.
E-mail: gmeghab@hotmail.com

This research is part of the ongoing study to better understand web page ranking on the web. It looks at a web page as a graph structure or a web graph, and tries to classify different web graphs in the new coordinate space: (out-degree, in-degree). The out-degree coordinate *od* is defined as the number of outgoing web pages from a given web page. The in-degree *id* 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 different web graphs are. *Google's* web ranking algorithm (Brin & Page, 1998) on ranking web pages is applied in this new coordinate space. The results of the algorithm has been modified to fit different topological web graph structures. Also the algorithm was not successful in the case of general web graphs and new ranking web algorithms have to be considered. This study does not look at enhancing web ranking by adding any contextual information. It only considers web links as a source to web page ranking. The author believes that understanding the underlying web page as a graph will help design better ranking web algorithms, enhance retrieval and web performance, and recommends using graphs as a part of visual aid for browsing engine designers.

1. Introduction

Scientific citations have been studied for a long time. Citation analysis in bibliometrics (Egghe & Rousseau, 1990) is the science of studying citations, their structures, and the evolution of a specific domain of knowledge from its citations. Many information sciences journals, e.g., *JASIST* (Small, 1973; Small, 1986), have devoted issues and complete volumes to the exploration of such an area of knowledge. All these confirm that:

- 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 contextual amount of variability between two articles in a given journal is usually small. Human judgment of the technicality of an article keeps the quality of publication at a 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 in a given year a journal j receives from other journal articles after it has been published for the last 2 years. It becomes the in-degree 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 link in a web page is dynamic and bidirectional. It points to other links, and other links point to it. A web page gets updated, so as new links are discovered, new links are added. It is a living community of links. The contextual amount of variability between two web pages on two separate web servers is very high. Not only do 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 rather than commonly established scientific explorations. Human judgment on web content is more subjective, and noisier than in a citation. No control of quality can be maintained on the web. Also, web pages can also be there merely for navigational purposes and not to link two documents. Page ranking is important because it makes use of the link structure of the web to calculate a quality ranking for each web page. Web pages are ranked according to a ranking scheme where each is evaluated according to the in-degree and the out-degree of that web page. Web pages also reflect the "signatures of intelligence" in our era and contain rich information on our collective society as a whole.

Received October 6, 2000; revised January 2, 2001; accepted January 11, 2001

The comparison above can seem to help or hinder the usage of the web link as a means to search the web. But the WWW and the scientific literature are governed by different principles. Web pages are not scrutinized but scientific journals are. New links in a web page can have an impact on the whole search done, and the ever-changing size of the web can reflect patterns of indexing, crawling for search engine (SE) designers who feed on these web links (Meghabghab, in press, a). WWW users will be more successful if they are made aware of how to interpret their search results when they query the WWW. 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, in press, b). A formula on how to measure the goodness of an SE emerges as follows from this study on the set of queries (SQ) used:

$$G = (SE; SQ) = \alpha^* \text{ (Coverage)} + \beta^* \text{ (Age)} \qquad (1)$$

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 SE designers.

To fully comprehend and assess the ranking and the influence of a web page over other web pages in a given web site, graph theory can be used to better understand the measures developed in Brin and Page (1998) and Brin (1998). Graph theory can also be used to discover new patterns that appear in a citation graph. The same idea can be used to measure the distance between two web pages. Discovering new ideas in a web page graph is easier than in a citation graph. Measuring the topological structure richness of a collection of web pages is an important aspect of web pages never explored before, and it is helpful in understanding page-ranking algorithms.

The next section is a reminder of concepts borrowed from graph theory to help analyze links, 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 also represents 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 make clear the topology of a web page and, in general, of the web space. In the space of all web pages W, let A ε 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) $\varepsilon\Re$, where $\Re$ is the relation "point to." We can add the following observations on the relation $\Re$:

A. If every web page is related to itself, we say that $\Re$ is reflexive.

B. For all web pages X and Y that belong to A, if (X, Y) $\varepsilon\Re \Rightarrow$ (Y, X) $\varepsilon\Re$. Web pages X and Y in that case represent mutual endorsement. Then the relation is said to be symmetric.

C. For all web pages X and Y that belong to A, if (X, Y) $\varepsilon\Re \Rightarrow$ (Y, X) $\varepsilon\Re$. Web pages X and Y are linked in a unidirectional way. Then the relation is then said to be anti-symmetric.

D. For all web pages that belong to A, 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:

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

E. 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:

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

F. When two web pages X and Y point to a distinct 3rd web page Z, then we could say that the 2 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 $\Re$.

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

G. 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 WWW by Larson (1996) and Pitkow and Pirolli (1997). Small and Griffith (1974) used breadth-first search to compute the connected components of the unidirected 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 $\Re$:

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

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" because almost all web pages point to that web page. Other emerging complex patterns can model web pages that can be qualified survey web pages or "hub web pages" because they cite "authoritative web pages" (Kleinberg, 1998).

2.1. Adjacency Matrix Representation of a Web Graph

Consider a graph G (Fig. 1) that represents a real web page and its adjacency matrix A. An entry a_{pq} in A is defined by the following:

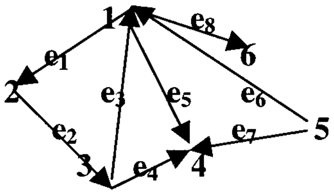

FIG. 1. A general web graph.

$$a_{pq} = 1 \quad \text{if there is an edge or link between 2 web pages p and q.}$$
$$= 0 \quad \text{Otherwise}$$

The Appendix expounds on the properties that can be discovered from an adjacency matrix.

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

$$a = A^T h \tag{2}$$

$$h = Aa \tag{3}$$

thus:

$$a = AA^T a \tag{4}$$

$$h = A^T Ah \tag{5}$$

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

- 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.
- An entry (p, p) in A^TA means the number of web pages that point towards p. We call that number the in-degree or id.
- An entry (p, q) in A^TA represents the number of web pages that are in common between p and q that point towards p and q.
- An entry (p, q) in AA^T represents the number of web pages that came out of p and q that are in common.

Here is the corresponding adjacency matrix for Figure 1:

$$A = \begin{array}{ccccccc} & & & & & & \text{od} \\ 0 & 1 & 0 & 1 & 0 & 1 & 3 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}$$

Building A^T yields:

$$A^T = \begin{array}{cccccc} 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{array}$$

Building C = AA^T yields:

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

Building D = A^TA yields:

$$D = A^TA = \begin{array}{cccccc} 2 & 0 & 0 & 2 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 3 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \end{array}$$

Figure 2 illustrates the in-degree and out-degree for the graph G.

How far away are two web pages in a web graph?

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, . . . A^n for a graph of n vertices. If we calculate the value of A^2 for Graph G, we have:

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

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 Figure 1 where (2,3) $\varepsilon\Re$ and (3,4) $\varepsilon\Re$.

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

Applying this to Graph G, one can see that $A^4 = A$. This means that the furthest away any two web pages are from each other is four web pages. An example in graph G is web

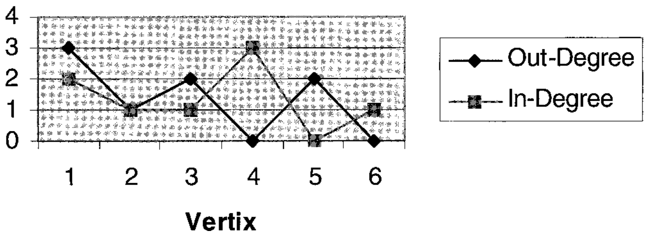

FIG. 2. In/out degrees for Figure 1.

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 would 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 Figure 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 as such:

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

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.

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

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 bidirectional links on this graph. That is why this graph is antisymmetric.

One way to look at how matrix A and vector id relate is by considering the following matrix-vector multiplication A^TU where A^T is the transpose of A already computed in 2.1 and U is the Unit vector 1.

Applying A^TU to Graph G results in the following:

$$A^T \times U = \begin{array}{cccccc} 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{array} \times \begin{array}{c} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{array} = \begin{array}{c} 2 \\ 1 \\ 1 \\ 3 \\ 0 \\ 1 \end{array}$$

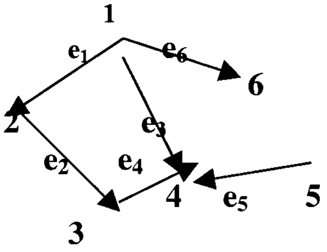

FIG. 3. A bipartite graph.

$$A^T \times U = id \tag{6}$$

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:

$$II^T = \begin{array}{cccccc} 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{array}$$

II^T is a diagonal matrix. Its eigenvalues vector is equal to vector id. Its eigenvectors constitute the columns of the unit matrix of size 6 × 6:

$$\text{Eigenvalue } (II^T) = id \tag{7}$$

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

$$\text{Eigenvalue } (II^T) = id = A^T \times U \tag{8}$$

2.3. 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 Figure 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 (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

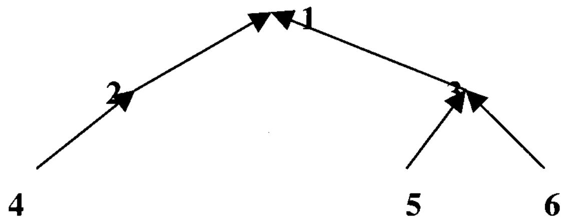

FIG. 4. An in-degree web graph.

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 so.

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

2.4. Web Topology

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

Different categories of web graphs can emerge according to their in-degree id or out-degree od. Even though we do not pretend that this classification is exhaustive, but we have gathered different kinds of graphs that were used to model different kinds of applications and domains of research. Emerging web graphs can be complex and rich in structure and links more than web page designers realize.

2.4.1. In-degree web graphs

Complex pages can emerge with large in-degree that looks like Figure 4.

Figure 5 illustrates such in-degree web pages by looking as their out/in-degree chart.

2.4.2. Out-degree web graphs

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

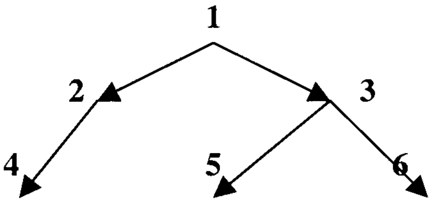

FIG. 6. An out-degree web graph.

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

2.4.3. Complete bipartite web graphs

Other complex web pages can emerge as complete bipartite graphs that look like Figure 8 with three nodes in the first set V_1 and three nodes in the second set V_2. (Note that the number of nodes in V_1 and V_2 is arbitrary.)

Remember that the topology of complete bipartite graphs like the one in Figure 8 is unique.

2.4.4. Bipartite web graphs

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

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 seen in Figure 10.

Pages with large in-degree or out-degree play an important role in web algorithms in general. The next two paragraphs illustrate that point.

2.5. Topological Difference of Web Pages

The signature of a web graph lies in its in-degree/out-degree characteristics, as can be seen from all these different charts. The in-degree/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 is the size of a graph important in the analysis of a graph, but its structure is also consequential. A graph will be assumed to be symbolized by two vectors: the in-degree vector and the out-degree vector. Next, the aver-

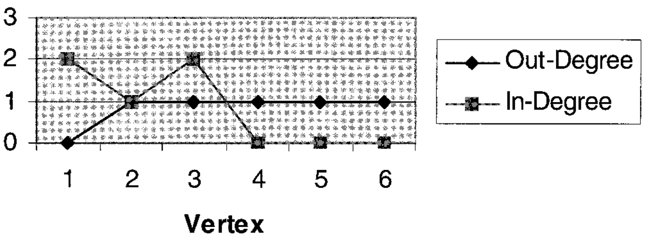

FIG. 5. Out/in degree of Figure 4.

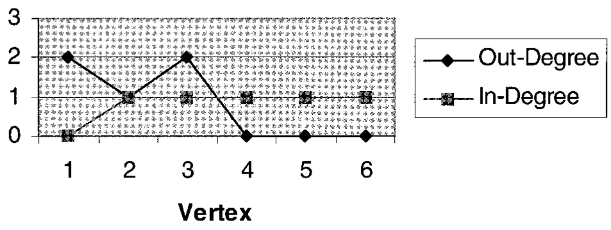

FIG. 7. Out/in degree of Figure 6.

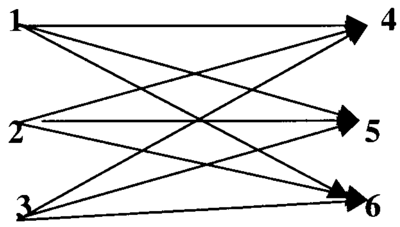

FIG. 8. A complete bipartite web graph.

age of the in-degree vector will be calculated, the average value of the out-degree vector will be calculated, and say that for a vertex v:

$$od(v) = \begin{array}{l} 1 \text{ if its value is above the average value} \\ 0 \text{ if otherwise.} \end{array} \qquad (9)$$

$$id(v) = \begin{array}{l} 1 \text{ if its value is above the average value} \\ 0 \text{ if otherwise.} \end{array} \qquad (10)$$

By applying (9) and (10) to Figure 1, which is a general graph, we could deduce:

$$od = (1,0,1,0,1,0)$$

$$id = (1,0,0,1,0,0)$$

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

$$od = (1,0,1,0,1,0)$$

$$id = (0,1,0,1,0,1)$$

By applying (9) and (10) to Figure 8, which is the only complete possible bipartite on figure 10, we deduce:

$$od = (1,1,1,0,0,0)$$

$$id = (0,0,0,1,1,1)$$

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

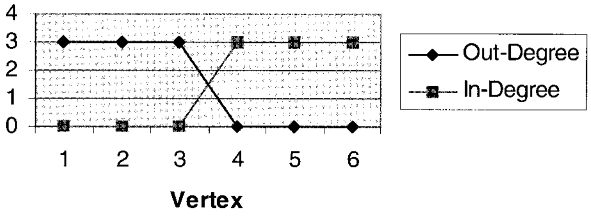

FIG. 9. Out/in degree of Figure 8.

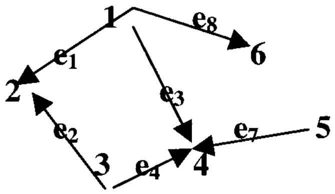

FIG. 10. A bipartite web graph.

$$od = (0,1,1,1,1,1)$$

$$id = (1,1,1,0,0,0)$$

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

$$od = (1,1,1,0,0,0)$$

$$id = (0,1,1,1,1,1)$$

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:

$$M = \begin{array}{c|ccccc} & 1 & 10 & 8 & 4 & 6 \\ 1 & (0,0) & (0,3) & (2,3) & (3,3) & (2,5) \\ 10 & (0,3) & (0,0) & (2,2) & (3,4) & (2,2) \\ 8 & (2,3) & (2,2) & (0,0) & (4,6) & (0,2) \\ 4 & (3,3) & (3,4) & (4,6) & (0,0) & (4,4) \\ 6 & (2,5) & (2,2) & (0,2) & (4,4) & (0,0) \end{array}$$

The smallest coordinate in this graph is the value (0,2), which says that Figures 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 coordinate in the graph is (0,3), which says that general graphs and bipartite graphs are the closest among all other graphs. The largest coordinate is (4,6), which says that complete bipartite graphs and in-degree are the farthest apart. The next biggest difference is between in-degree and out-degree trees, which is evident form 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 is can be concluded from the statement before.

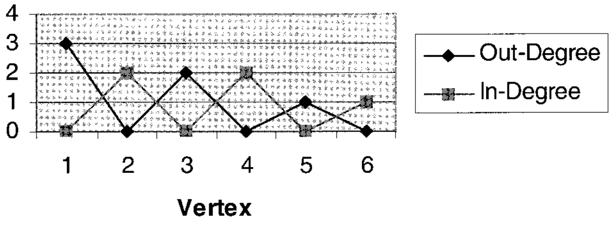

FIG. 11. Out/in degree of Figure 10.

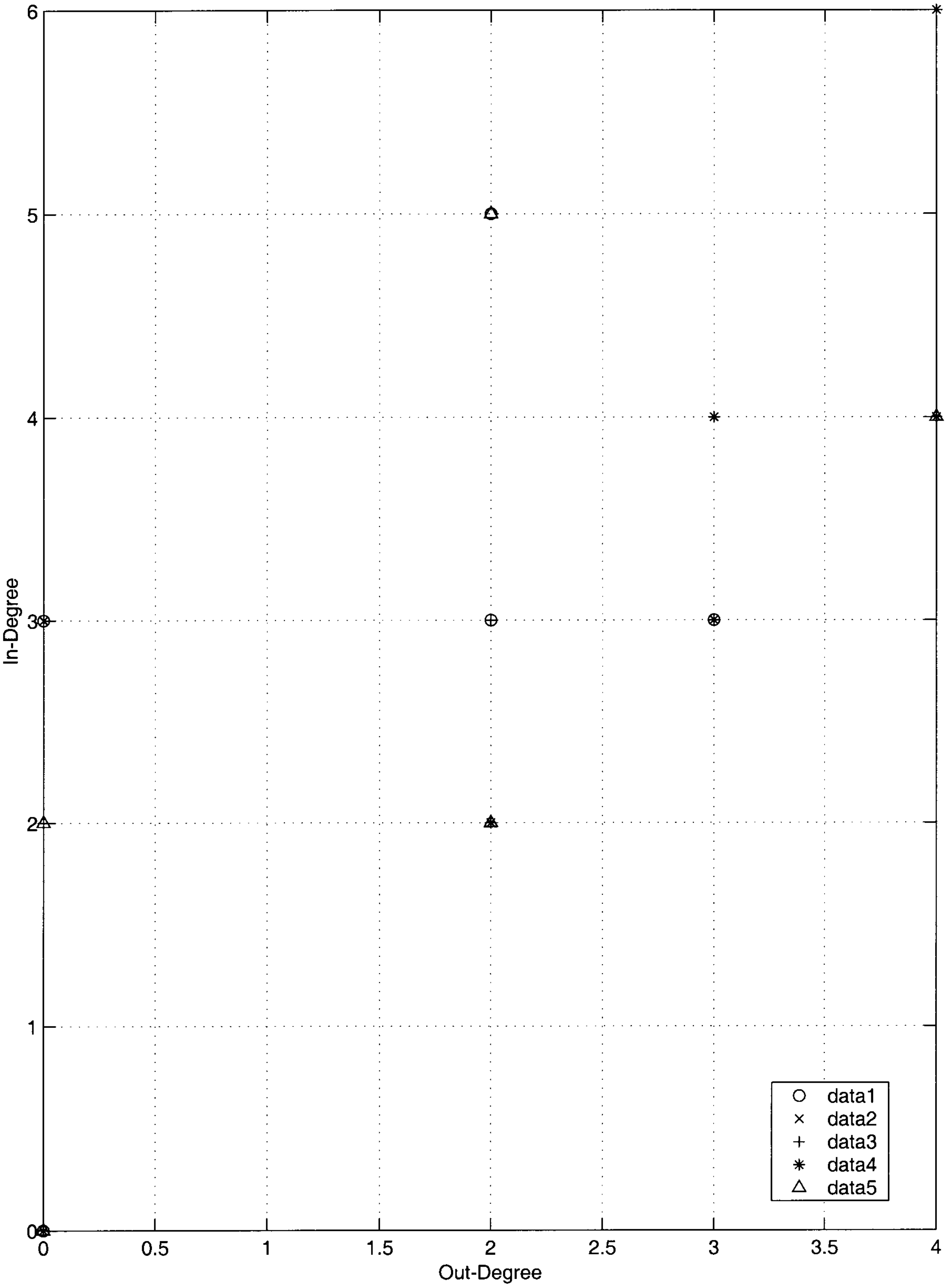

FIG. 12. Classification of web graphs and the new coordinate space.

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

$$|(\mathrm{CBG})| < |(\mathrm{OD})| < |(\mathrm{BG})| < |(\mathrm{GG})| < |(\mathrm{ID})| \quad (11)$$

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

Figure 12 displays the classification of the web graphs in the new coordinate (out-degree, in-degree) space. In Figure 12, data 1 represents all the data in the first row of matrix M, data 2 represents all the data in the second row of matrix M, data 3 represents all the data in the third row of matrix M, data 4 represents all the data in the fourth row of matrix M, and data 5 represents all the data in the fifth row of matrix M. Pages with large in-degree or out-degree play an important role in web algorithms in general. The next algorithm shows that point. No other studies have reflected on the topological structure of web graphs and web algorithms.

Page Ranking According to Google's Search Engine (Brin & Page, 1998)

Let r_p be the rank of a page p and od_p the out-degree of node i, i.e., the number of outgoing links from a web page p. The rank of a page p is computed as specified by Google's SE:

$$r_p = (1 - c) + c[(r_1/od_1) + (r_2/od_2) + (r_3/od_3) + \cdots + (r_n/od_n)] \quad (12)$$

for all nodes 1,2,3, ... n where (1,p) ((, (2,p) ((,(3,p) ((, (n,p) ((and where c is a constant: $0<c<1$ (ideally selected at 0.85).

Note that the r_p form a probability distribution over all web pages, so the probability over all web pages will be one.

Applying equation (12) to the graph in Figure 1 yields the following equations for all these vertices:

$$r_1 = (1 - c) + c(r_3/2 + r_5/2) \quad (13)$$
$$(3,1)\ \varepsilon\Re,(5,1)\ \varepsilon\Re$$

$$r_3 = (1 - c) + c(r_2) \quad (14)$$
$$(2,3)\ \varepsilon\Re$$

$$r_5 = (1 - c) + 0 = 1 - c \quad (15)$$

$$r_2 = (1 - c) + c(r_1/3) \quad (16)$$
$$(1,2)\ \varepsilon\Re$$

$$r_4 = (1 - c) + c(r_1/3 + r_3/2 + r_5/2) \quad (17)$$
$$(1,4)\ \varepsilon\Re,\ (3,4)\ \varepsilon\Re,(5,4)\ \varepsilon\Re$$

$$r_6 = (1 - c) + c(r_1/3) \quad (18)$$
$$(1,6)\ \varepsilon\Re$$

Replacing (14), (15), (16) in (13) yields:

$$r_1 = 3(1 - c)(c + 3 + c^2)/(6 - c^3)$$

Table 1 shows the values of r_1, r_2, r_3, r_4, r_5, and r_6 as a function of c.

TABLE 1. Summary of Google's ranking applied to Figure 13.

Rank of vertex	Value of c	Value of id
r_1 = 0.38	0.85	2
r_2 = 0.26	0.85	1
r_3 = 0.37	0.85	1
r_4 = 0.43	0.85	3
r_5 = 0.15	0.85	0
r_6 = 0.26	0.85	1

The following observations can be made:

Vertex 4 being the vertex with the highest links pointing to it or degree id = 3 yielded the highest ranking among all the pages.

Vertex 1 follows vertex 4 in its ranking because of id = 2.

Vertices 3, 2, and 6 rank behind vertex 1. Although these three vertices have their id = 1, they should have equal value r; yet they differ greatly. Value r_3 is closer to r_1 than it is to the rank of vertices 2 and 6.

Vertex 5 with the least rank because it did not have any nodes pointing to it.

Vertices with the same in-degree id according to equation (12) should yield equal ranking. That last principle is violated in graph G. Vertices 2, 3, and 6 should have very similar ranking. To observe whether Google's ranking behaves well on other topological structures, we simulated Google's algorithm on more complex patterns, i.e., in-degree web graphs like Figure 4, out-degree web graphs like Figure 6, bipartite graphs like Figure 3, and complete bipartite graphs like Figure 8.

Theorem 1: Nodes at the same level in complete bipartite graphs will have equal value to Google's rank.

$$r_0 = 0.15$$

$$r_1 = 0.15 + 0.85(r_0 * m/k) \quad (19)$$

where m is the number of nodes at level = 0, and k is the number of children to each node in level = 0

Proof: Given the fact complete Bipartite graphs are made of two levels, one level with id = 0 and another level = 1 with id higher than 0, then all vertices of a degree higher than 0 will have equal ranking regardless of the number of nodes or configurations.

Example: In Figure 13, m = 4 and k = 3. The following observations can be made:

- All vertices in V_1 = {1,2,3,4} have an id = 0. They are also at level = 0. Their ranks according to Theorem 1 will be all the same and will equal to:

$$r_0 = 0.15(\text{for } c = 0.85)$$

- All vertices in V_2 = {5,6,7} have an id = 4. They are all at level = 1. Their ranks according to Theorem 1 will be the same and will be equal to:

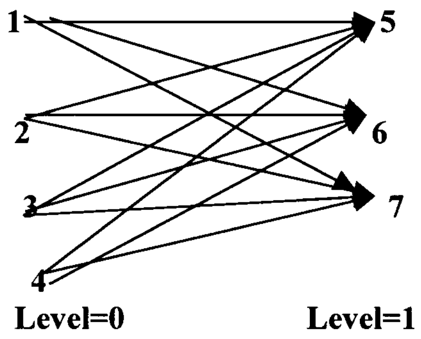

FIG. 13. A complete bipartite web graph.

$$r_1 = (1 - c) + c(r_0*4/3) = 0.15 + 0.85(0.15*4/3) = 0.32$$

Theorem 2: Nodes at the same level in out-degree trees will have equal value to Google's rank.

$$r_n = 0.15 \quad \text{for } n = 0$$

$$r_n = 0.15 + 0.85*(r_{n-1}/m) \quad \text{for } n = 1 \qquad (20)$$

$$\text{One Sub-tree: } r_n = 0.15 + 0.85*(r_{n-1}/m) \quad \text{for } n > 1 \text{ and if } m\#k$$

$$\text{Other Sub-tree: } r_n = 0.15 + 0.85*(r_{n-1}/k)$$

where m is the number of children of node n − 1 in one subtree, k is the number of children of node n − 1 in the other subtree, and L is the level of the node.

Proof: Out-degree graphs (or trees) having more than two levels will never provide equal ranking for vertices with equal degree. Instead of the degree of a node, the level of a node in a given tree will be the determining factor.

Example: Consider the following out-degree tree in Figure 14.

$$r_0 = 0.15$$

$$r_1 = 0.15 + 0.85*(r_0/2) = 0.15 + 0.85*0.075 = 0.214$$

for n = 2, we have 2 ranks because $k\#m$ ($m = 3$, $k = 2$):

$$r_2 = 0.15 + 0.85(r_1/3) = 0.15 + 0.85*(0.214/3) = 0.221$$

$$r_2 = 0.15 + 0.85(r_1/2) = 0.15 + 0.85*(0.214/2) = 0.241$$

Thus the rank of node 1 is r_0, the rank of nodes 2 and 3 is the same r_1, the rank of nodes 4, 5, 6 is the same r_2 = 0.221, and the rank of nodes 7 and 8 is the same r_2 = 0.241.

Interpretation: Even though all vertices are of the same degree id = 1 except the root of the tree, which has a degree id = 0, vertices 2, and 3 at level = 1 will have equal rank, which is 0.214. Vertices at level = 2 will be divided into two ranks because the left subtree has more children coming out of node 2 than those coming out of node 3 on the right subtree. Vertices 4, 5, 6, which are at level 2 of the let subtree, will have lower rank than those on the right part of the tree, which are still higher than that of those vertices at level = 1. This is expected, since what feeds into vertices 4, 5, 6 are nodes at a degree higher than those feeding into vertices 2 and 3, which is node 1. If 1 cites 2 and 3, and 2 now cites three others like 4, 5, and 6, then 4, 5, and 6 will have higher ranking than 2 because they are more important than 2.

Theorem 3: Nodes at the same level in in-degree trees will have equal value to Google's rank.

$$r_0 = 0.15$$

$$r_n = 0.15 + 0.85*(r_{n-1} * m + r_{n-2} * k) \qquad (21)$$

where m is the number of parents to node n from level n − 1 and k is the number of parents to node n from level n − 2.

Proof: In-degree graphs (or trees) having more than two levels will never provide equal ranking for vertices with equal in-degree. Instead of the in-degree of a node, the level of a node in a given tree will be the determining factor.

Example: Consider the following general in-degree tree in Figure 15.

$$r_0 = 0.15$$

$$r_1 = 0.15 + 0.85*(r_0*2) = 0.15 + 0.85*0.30 = 0.405$$

$$r_2 = 0.15 + 0.85(r_1*2 + 1*r_0) = 0.15 + 0.85*(0.405 + 0.15) = 0.555$$

Thus the rank of node 8 is r_2, the rank of nodes 6 and 7 is r_1, and the rank of nodes 1, 2, 3, 4, and 5 is r_0.

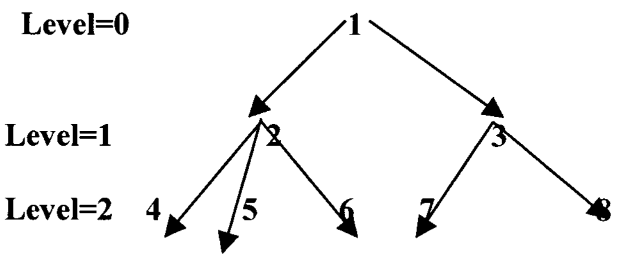

FIG. 14. An out-degree web graph.

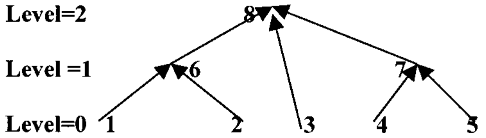

FIG. 15. An in-degree web graph.

Interpretation: Even though vertices 6, 7, and 8 have the same degree id = 2, only vertices 6, and 7 at level = 1 will have equal rank, which is 0.405. Vertex 8, which is at level 2, will have the highest rank that is 0.555. Vertices 1, 2, 3, 4, and 5, which are at level 0, will have equal rank, which is much lower than that of all other vertices. This is expected since what feeds into vertex 8 are not only a low level node such as 3, but also vertices 6 and 7, which are nodes at an in-degree higher than those feeding into 6 and 7, which are vertices 1, 2, 4, and 5, which occupy level = 0. If 1 and 2 point to 6, 4 and 5 point to 7, 6 and 7 point to 8 besides 3, that makes 8 a more a more important link than all other links in the tree.

Theorem 4: Google's ranking procedure will not work well on bipartite graphs.

Proof: Even though we have seen that bipartite graphs allow the separation of the set of vertices into two sets, nothing has changed in the interpretation of the fact that nodes with equal in-degree id can have different rankings, which is contrary to the interpretation of equation (12).

Example: Consider the example of a bipartite graph like the one in Figure 3. Table 2 summarizes Google's ranking applied to Figure 3.

The same remark that was applied to the graph in Figure 1 still applies to the graph in Figure 3. Vertex 3 has a higher ranking than those of the 3 vertices 2, 3, and 6, which have the same in-degree. Making a graph a bipartite graph did not change anything in the ranking of links. A more involved ranking scheme will be needed to further explore the complexity of ranking in bipartite and general graphs.

Theorem 5: Google's ranking procedure will not work well on general graphs.

Table 3 summarizes the results of the modified Google's ranking procedure on these different topological structures. Thus, such an algorithm needs to be applied carefully depending upon the topological structure of the web pages that are studied. No other studies have reported the influence of the topology of the web links on Google's page-ranking technique.

TABLE 2. Summary of Google's ranking applied to Figure 3.

Rank of vertex	Value of c	Value of id
$r_1 = 0.15$	0.85	0
$r_2 = 0.21$	0.85	1
$r_3 = 0.33$	0.85	1
$r_4 = 0.62$	0.85	3
$r_5 = 0.15$	0.85	0
$r_6 = 0.21$	0.85	1

TABLE 3. Summary of Google's ranking applied to different web graphs.

Topological structure	Google's web ranking
In-degree trees	Works well per tree level (see Theorem 3)
Out-degree trees	Works well per subtree level (see Theorem 2)
Bipartite graphs	Does not work (see Theorem 4)
General graphs	Does not work (see Table 1 and Theorem 5)
Complete bipartite graphs	Works well per graph level (see Theorem 1)

Final Observations on Google's Ranking Algorithm and Conclusion

Google's ranking algorithm needs to be adjusted to different topological web structures to be able to successfully rank web pages without any contextual information added. Given the fact that Google's SE is gaining momentum in indexing more than one billion web pages and being adopted by major SEs like Yahoo, it seems that a study of their ranking web algorithm is timely in further exploring its applicability in a variety of web graphs. This study focused first on categorizing different web graphs and how close or far away these different topological web graphs are. Then we applied Google's web ranking algorithm to the complete bipartite graphs, followed by bipartite graphs, then out-degree trees, in-degree trees, and lastly, general graphs. Google's ranking web algorithm worked best on complete bipartite graphs by ranking equally vertices at the same level. The algorithm did not fare well in other web graph structures on the lower ranking of the remaining vertices. More specifically, vertices with equal degrees (e.g., equal amount of outgoing nodes) did not rank equally. Different theorems were adopted for these different topological structures, and Google's was readjusted ranking to fit these different topological structures.

The metric adopted here to classify these different structures in the coordinate (out-degree, in-degree) space has been applied in a more recent study (Meghabghab, in press, a) to discovering hubs and authorities in a variety of graph web-based pages. Early results show the uncommon phenomenon that a web page can be both a hub page and an authority web page in general graphs only, for example. The other graphs showed that phenomenon in the beginning of the filtering procedure, but after a number of iterations, the only web pages left were those that were either hub web pages or authority web pages, and not both at the same time. A web page is said to be an authority web page (Kleinberg, 1998) if many web pages point to it. The web pages that point to those authority web pages are themselves called hub web pages. According to Kleinberg (personal e-mail to the author, Sept. 14, 2000), that phenomenon is quite uncommon even though the web algorithm itself does not prohibit it from happening. 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 and hence the performance of retrieval heuristics. In recent studies, Bharat and Henzinger (1998), Chakrabarti et al. (1998a), and Charkrabarti et al. (1998b) performed three user studies to evaluate the HITS system to better find information 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. (1998), 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; 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.

Other ranking algorithms could have been studied and applied to the different topological web graphs, but given the popularity of Google and its wide indexing power, with more than a billion web pages, makes it a very powerful SE that has been adopted by other SEs like Yahoo. All of these factors contributed to the consideration of Google’s ranking algorithm in this study. The author is considering reviewing other web page-ranking algorithms to be applied to the same rich topological web graphs.

Appendix

Consider a graph G that represents a real web page and its adjacency matrix A. An entry a_{pq} in A is defined by the following:

$a_{pq} = 1$ if there is an edge or link between 2 web pages p and q.
$= 0$ Otherwise

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

A. 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.

B. A graph is said to be symmetric if for all edges p and q in G: iff $a_{pq} = 1$ then $a_{qp} = 1$. We say in this case that there is mutual endorsement.

C. 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.

D. A graph is said to be transitive if for all edges p,q, and r:

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

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

E. A graph is said to be antisymmetric iff for all edges p and q:

$$Iff\ a_{pq} = 1 \quad \text{then} \quad a_{qp} = 0$$

F. 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.

$$a_{pr} = 1 \quad \text{and} \quad a_{qr} = 1$$

G. If a single page p points to two different web pages q and r then we say that there is co-citation.

$$a_{pq} = 1 \quad \text{and} \quad a_{pr} = 1$$

To illustrate the above points let us look back again at Figure 1.

Here are the algebraic properties of $\Re$ in G:

$\Re$ is not reflexive
$\Re$ is not symmetric
$\Re$ is not transitive
$\Re$ is anti-symmetric
(1,4) $\varepsilon\Re$, (3,4) $\varepsilon\Re$, and (5,4) $\varepsilon\Re$:
we could say that the vertex with the highest number of web pages pointing to it.
(5,1) $\varepsilon\Re$ and (5,4) $\varepsilon\Re$: 5 co-cites 1 and 4.

Acknowledgment

Special thanks are given to *JASIST* referees for thorough review of this paper and valuable feedback.

References

Albert, R., Jeong, H., & Barabasi, A.L. (1999). Diameter of the World Wide Web. Nature, 401, 130–131.

Bharat, K., & Henzinger, M.R. (1998). Improved algorithms for topic distillation in a hyper-linked environment. Proceedings of the ACM Conference on Research and Development in Information Retrieval.

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

Brin, S., & Page, L. (April 1998). The anatomy of a large scale hypertextual web search engine. Proceedings of the 7th 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. Proceedings of the 7th 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. ACM SIGIR Workshop on Hypertext Information Retrieval on the Web.

Egghe, L., & Rousseau, R. (1990). Introduction to infometrics. Elsevier.

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

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

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

Kleinberg, J. (January 1998). Authoritative sources in a hyper-linked environment. Proceedings of the ACM-SIAM Symposium on Discrete Algorithms (pp. 668–677).

Larson, R. (1996). Bibliometrics of the World Wide Web: An exploratory analysis of the intellectual structure of cyberspace. Annual Meeting of the American Society for Information Science.

McBryan, O. (1994). GENVL and WWW: Tools for taming the Web. Proceedings of the 1st International World Wide Web Conference.

Meghabghab, G. (2000). Stochastic Simulations of Rejected World Wide Web Pages. Proceedings of the 8th 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. (2001). 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(4), 571–591.

Meghabghab, G. (in press) Discovering authorities and hubs in different topological web graph structures. Information Processing and Management.

Mukherjea, S., Foley, J., & Hudson, S. (1995). Visualizing complex hypermedia networks through multiple hierarchical views. Proceedings of ACM SIGCHI Conference in Human Factors in Computing (pp. 331–337). Denver, Colorado.

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. Proceedings of ACM SIGCHI Conference in Human Factors in Computing.

Pitkow, J., & Pirolli, P. (1997). Life, death, and lawfulness on the electronic frontier. 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 Science, 24, 265–269.

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

Small, H., & Griffith, B.C. (1974) The structure of the scientific literatures. I. Identifying and graphing specialties. Science Studies, 4, 17–40.

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