A Survey of Web Metrics

DEVANSHU DHYANI, WEE KEONG NG, AND SOURAV S. BHOWMICK

Nanyang Technological University

The unabated growth and increasing significance of the World Wide Web has resulted in a flurry of research activity to improve its capacity for serving information more effectively. But at the heart of these efforts lie implicit assumptions about "quality" and "usefulness" of Web resources and services. This observation points towards measurements and models that quantify various attributes of web sites. The science of measuring all aspects of information, especially its storage and retrieval or *informetrics* has interested information scientists for decades before the existence of the Web. Is Web informetrics any different, or is it just an application of classical informetrics to a new medium? In this article, we examine this issue by classifying and discussing a wide ranging set of Web metrics. We present the origins, measurement functions, formulations and comparisons of well-known Web metrics for quantifying *Web graph properties*, *Web page significance*, *Web page similarity*, *search and retrieval*, *usage characterization* and *information theoretic properties*. We also discuss how these metrics can be applied for improving Web information access and use.

Categories and Subject Descriptors: H.1.0 **[Models and Principles]**: General; H.3.3 **[Information Storage and Retrieval]**: Information Search and Retrieval; I.7.0 **[Document and Text Processing]**: General

General Terms: Measurement

Additional Key Words and Phrases: Information theoretic, PageRank, quality metrics, Web graph, Web metrics, Web page similarity

1. INTRODUCTION

The importance of measuring attributes of known objects in precise quantitative terms has long been recognized as crucial for enhancing our understanding of our environment. This notion has been aptly summarized by Lord Kelvin:

> "When you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot express it in numbers, your knowledge is of a meager and unsatisfactory kind; it may be the beginning of knowledge, but you have scarcely in your thoughts advanced to the state of science."

One of the earliest attempts to make global measurements about the Web was undertaken by Bray [1996]. The study attempts to answer simple questions on

Authors' address: College of Engineering, School of Computing Engineering, Nanyang Technological University, Blk N4-2A-32, 50 Nanyang Avenue, Singapore 639789, Singapore; email: {assourav,awkng}@ntu.edu.sg.

attributes such as the size of the Web, its connectivity, visibility of sites and the distribution of formats. Since then, several directly observable metrics such as hit counts, click-through rates, access distributions and so on have become popular for quantifying the usage of web sites. However, many of these metrics tend to be simplistic about the phenomena that influence the attributes they observe. For instance, Pitkow [1997] points out the problems with hit metering as a reliable usage metric caused by proxy and client caches. Given the organic growth of the Web, we require new metrics that provide deeper insight on the Web as a whole and also on individual sites from different perspectives. Arguably, the most important motivation for deriving such metrics is the role they can play in improving the quality of information available on the Web.

To clarify the exact meaning of frequently used terms, we supply the following definition [Boyce et al. 1994]:

> Measurement, in most general terms, can be regarded as the assignment of numbers to objects (or events or situations) in accord with some rule [*measurement function*]. The property of the objects that determines the assignment according to that rule is called *magnitude*, the measurable attribute; the number assigned to a particular object is called its *measure*, the amount or degree of its magnitude. It is to be noted that the rule defines both the magnitude and the measure.

In this article, we provide a survey of well-known metrics for the Web with regard to their magnitudes and measurement functions. Based on the attributes they measure, these are classified into the following categories:

—*Web Graph Properties*. The World Wide Web can be represented as a graph structure where Web pages comprise nodes and hyperlinks denote directed edges. Graph-based metrics quantify structural properties of the Web on both macroscopic and microscopic scales.

—*Web Page Significance*. Significance metrics formalize the notions of "quality" and "relevance" of Web pages with respect to information needs of users. Significance metrics are employed to rate candidate pages in response to a search query and have an impact on the quality of search and retrieval on the Web.

—*Usage Characterization*. Patterns and regularities in the way users browse Web resources can provide invaluable clues for improving the content, organization and presentation of Web sites. Usage characterization metrics measure user behavior for this purpose.

—*Web Page Similarity*. Similarity metrics quantify the extent of relatedness between web pages. There has been considerable investigation into what ought to be regarded as indicators of a relationship between pages. We survey metrics based on different concepts as well as those that aggregate various indicators.

—*Web Page Search and Retrieval*. These are metrics for evaluating and comparing the performance of Web search and retrieval services.

—*Information Theoretic*. Information theoretic metrics capture properties related to information needs, production and consumption. We consider the relationships between a number of regularities observed in information generation on the Web.

We find that some of these metrics originate from diverse areas such as classical informetrics, library science, information retrieval, sociology, hypertext and econometrics. Others, such as Web-page-quality metrics, are entirely specific to the Web. Figure 1 shows a taxonomy of the metrics we discuss here.

Metrics, especially those measuring phenomena, are invariably proposed in the context of techniques for improving the quality and usefulness of measurable objects; in this case, information on the Web. As such, we also provide some insight on the applicability of Web metrics. However, one must understand that their usefulness is limited by the models that explain the underlying phenomena

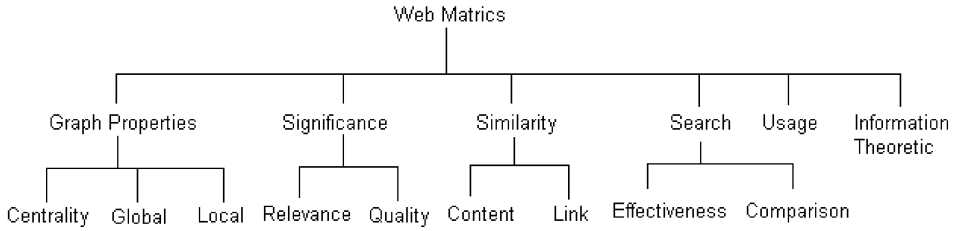

Fig. 1. A taxonomy of Web metrics.

and establish causal relationships. A study of these metrics is a starting point for developing these models, which can eventually aid Web content providers in enhancing web sites and predicting the consequences of changes in certain attributes.

2. WEB GRAPH PROPERTIES

Web graph properties are measured by considering the Web or a portion of it, such as a web site, as a directed hypertext graph where nodes represent pages and edges hyperlinks referred to as the *Web graph*. Web graph properties reflect the structural organization of the hypertext and hence determine the readability and ease of navigation. Poorly organized web sites often cause user disorientation leading to the "lost in cyberspace" problem. These metrics can aid web site authoring and create sites that are easier to traverse. Variations of this model may label the edges with weights denoting, for example, connection quality, or number of hyperlinks. To perform analysis at a higher granular level, nodes may be employed to model entire web sites and edges after the total strength of connectivities amongst web sites. In the classification below, we first consider the simplest hypertext graph model to discuss three types of graph properties introduced by Botafogo et al. [1992]; namely, *centrality*, *global measures* and *local measures*. Then, we discuss *random graph model* based on random networks.

Before discussing metrics for hypertext graph properties, we introduce some of the preliminary terms. The hypertext graph of N nodes (Web pages) can be represented as an $N \times N$ *distance matrix*[1] C, where element C_{ij} is the number of links that have to be followed to reach node j starting from node i or simply the distance of j from i. If nodes i and j are unconnected in the graph, C_{ij} is set to a suitable predefined constant K. Figure 2 shows an example hyperlink graph. The distance matrix for this graph is shown in Table I.

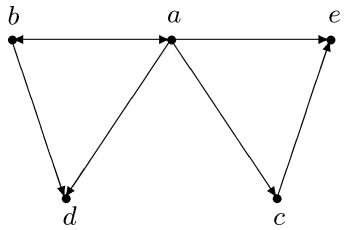

Fig. 2. Hyperlink graph example.

Table I. Distance Matrix and Associated Centrality Metrics; $K = 5$

Nodes	a	b	c	d	e	OD	ROC
a	0	1	1	1	1	4	17
b	1	0	2	2	3	8	8.5
c	5	5	0	5	1	16	4.25
d	5	5	5	0	5	20	3.4
e	5	5	5	5	0	20	3.4
ID	16	16	13	13	10	68	
RIC	4.25	4.25	5.23	5.23	6.8		

2.1. Centrality

Centrality measures reflect the extent of connectedness of a node with respect to other nodes in the graph. They can be used to define hierarchies in the hypertext with the most central node as the root. The *out distance* (*OD*) of a node i is defined as the

[1] We differ slightly from Botafogo et al. [1992], where this matrix is referred to as the *converted distance matrix*.

sum of distances to all other nodes; that is, the sum of all entries in row i of the distance matrix C. Similarly, the *in distance* (ID) is the sum of all in distances. Formally,

$$OD_i = \sum_j C_{ij}$$
$$ID_i = \sum_j C_{ji}.$$

In order to make the above metrics independent of the size of the hypertext graph, they are normalized by the *converted distance* or the sum of all pair-wise distances between nodes thereby yielding the *relative out centrality* (ROC) and the *relative in centrality* (RIC) measures, respectively. Therefore,

$$ROC_i = \frac{\sum_i \sum_j C_{ij}}{\sum_j C_{ij}}$$
$$RIC_i = \frac{\sum_i \sum_j C_{ij}}{\sum_j C_{ij}}.$$

The calculation of centrality metrics for the graph in Figure 2 is shown in Table I. A central node is one with high values of relative in- or out-centrality suggesting that the node is close to other nodes in hyperspace. Identification of a root node (a central node with high relative out-centrality) is the first step towards constructing easily navigable hypertext hierarchies. The hypertext graph may then be converted to a crosslinked tree structure using a breadth-first spanning tree algorithm to distinguish between hierarchical and cross-reference links.

2.2. Global Metrics

Global metrics are concerned with the hypertext as a whole and not individual nodes. They are defined in a hierarchically organized hypertext where the hierarchical and cross-referencing links are distinguished. Two global metrics discussed here are the *compactness* and *stratum*. The *compactness* metric indicates the extent of cross referencing; a high compactness means that each node can easily reach other nodes in the hypertext. Compactness varies between 0 and 1; a completely disconnected graph has compactness 0 while a fully connected graph has compactness 1. For high readability and navigation, both extremes of compactness values should be avoided. More formally,

$$C_p = \frac{\text{Max} - \sum_i \sum_j C_{ij}}{\text{Max} - \text{Min}},$$

where Max and Min are, respectively, the maximum and minimum values of the centrality normalization factor—converted distance. It can be shown that Max and Min correspond to $(N^2 - N)K$ (for a disconnected graph) and $(N^2 - N)$ (for a fully connected graph), respectively. We note that the compactness $C_p = 0$ for a disconnected graph as the converted distance becomes Max and $C_p = 1$ when the converted distance equals Min for a fully connected graph.

The *stratum* metric captures the linear ordering of the Web graph. The concept of stratum characterizes the linearity in the structure of a Web graph. Highly linear web sites, despite their simplicity in structure, are often tedious to browse. The higher the stratum, the more linear the Web graph in question. Stratum is defined in terms of a sociometric measure called *prestige*. The prestige of a node i is the difference between its status, the sum of distances *to* all other nodes (or the sum of row i of the distance matrix), and its contrastatus, the sum of finite distances *from* all other nodes (or the sum of column i of the distance matrix). The *absolute prestige* is the sum of absolute values of prestige for all nodes in the graph. The stratum of the hypertext is defined as the ratio of its absolute prestige to *linear absolute prestige* (the prestige of a linear hypertext with equal number of nodes). This normalization by linear absolute prestige renders the prestige value insensitive to the hypertext size. Formally,

$$S = \frac{\sum_i (\mid \sum_j C_{ij} - \sum_j C_{ji} \mid)}{\text{LAP}},$$

Table II. Distance Matrix and Stratum Related Metrics

Nodes	a	b	c	d	e	Status	Absolute prestige
a	0	1	1	1	1	4	3
b	1	0	2	2	3	8	7
c	∞	∞	0	∞	1	1	2
d	∞	∞	∞	0	∞	0	3
e	∞	∞	∞	∞	0	0	5
Contrastatus	1	1	3	3	5	13	20

where the linear absolute prestige LAP is the following function of the number of nodes N.

$$\text{LAP} = \begin{cases} \dfrac{N^3}{4}, & \text{if } n \text{ is even} \\ \dfrac{N^3 - N}{4}, & \text{otherwise.} \end{cases}$$

Stratum metrics for the graph of Figure 2 are shown in Table II. The computation of stratum only considers *finite* distances; hence, we invalidate unconnected entries in the distance matrix (denoted ∞). The linear absolute prestige for a graph of 5 nodes from the above formula is 30. The stratum of the graph can be calculated by normalizing the sum of absolute prestige (Table II) of all nodes by the LAP. For the graph of Figure 2, this equals 0.67.

We conclude the survey of global metrics by citing some measurements of *degree distributions* that are reported in Kleinberg et al. [1999] and Kumar et al. [1999] and confirmed in Broder et al. [2000]. In experiments conducted on a subgraph of the Web, the in-degree distribution (in-degree versus frequency) has been found to follow *Lotka's law*. That is, the probability that a node has in-degree i is proportional to $1/i^{\alpha}$, where α is approximately 2. A similar observation holds for the out-degree distribution. In Dhyani [2001], we describe our own experiments to confirm these findings and use the observation as the premise for ascertaining distributions of some well-known hyperlink-based metrics and, subsequently, we derive other important informetric laws related to Lotka's law.

Attempts have also been made to study the macroscopic structure of the WWW. In their experiments on a crawl of over 200 million pages, Broder et al. [2000] found that over 90% percent of the Web comprises a single connected component[2] if the links are treated as undirected edges. Of these, a core of approximately 56 million forms a strongly connected component. The maximum distance between any two pages or the *diameter* of this core is only 28 as compared to a diameter of over 500 for the entire Web graph. The probability that a path exists between two randomly chosen pages was measured to be 24%. The average directed path length is 16.

The significance of the above observations are two-fold. First, they become the starting point for modelling the graph structure of the Web. For example, Kleinberg et al. [1999] have explained the degree distributions by modeling the process of copying links while creating Web pages. These models can be of use in predicting the behavior of algorithms on the Web and discovering other structural properties not evident from direct observation. Secondly, knowledge of the structure of the Web and its graph properties can lead to improved quality of Web search as demonstrated by hyper link metrics such as PageRank [Brin and Page 1998], Authorities/Hubs [Kleinberg 1998], and Hyperinformation [Marchiori 1997].

2.3. Local Metrics

Local metrics measure characteristics of individual nodes in the hypertext graph. We discuss two local metrics; namely, *depth* and *imbalance* [Botafogo et al. 1992]. The *depth* of a node is just its distance from the root. It indicates the ease with which the node in question can be reached and consequently its importance to the reader. That is, the bigger the distance of a node from the root, the harder it is for the reader to reach this

[2] A portion of the Web graph such that there exists a path between any pair of pages.

node, and consequently the less important this node will be in the hypertext. Nodes that are very deep inside the hypertext are unlikely to be read by the majority of the readers [Botafogo et al. 1992]. Note that an author may intentionally store a low relevance piece of information deep inside the hypertext. Consequently, the readers whose interest is not so strong can browse the hypertext without seeing the low relevance information, while more interested readers will be able to have a deeper understanding of the subject by probing deeper into the hypertext. Having access to a depth metric, web-site designers can locate deep nodes and verify that they were intentional.

The *imbalance* metric is based on the assumption that each node in the hypertext contains only one idea and the link emanating from a node are a further development on that idea (except cross-reference links). Consequently, we might want the hypertext to be a balanced tree. The *imbalance* metric identifies nodes that are at the root of imbalanced trees and enables the web-site designer to identify imbalance nodes. Observe that imbalance in a hypertext does not necessarily indicate poor design of hypertext [Botafogo et al. 1992]. The crux of the matter is that each topic should be fully developed. Thus, in a university department hypertext, there may be many more levels of information on academic staffs and their research areas than on locations of copying machines. However, we should expect that each of the major areas in the hypertext is treated with equal importance. Although balance is not mandatory, too much imbalance might indicate bias of the designer or a poorly designed web site [Botafogo et al. 1992]. Note that, similar to depth the imbalance metric can be used as a feedback to the authors. If they decide that imbalances are desired, then they can overlook the information.

In order to quantify imbalance, Botafogo et al. [1992] proposed two imbalance metrics: *absolute depth imbalance* and *absolute child imbalance*. Let T be a general rooted tree. Let $a_1, a_2, \ldots, a_n$ be children of node a. Then, the *depth vector* $D(a)$ [Botafogo et al. 1992] is defined as follows:

$$D(a) = \begin{cases} [1 + Max(D(a_1)), 1 + Max(D(a_2)), \ldots, 1 + Max(D(a_n))], \\ [0] \quad \text{if } a \text{ has no child } (n = 0), \end{cases}$$

where $Max(D(a_i))$ indicates the value of the largest element in the vector $D(a_i)$. The depth vector is represented within square brackets. Intuitively, this vector indicates the maximum distance one can go by following each of the children of node a. The *absolute depth imbalance* of a node a is defined as the standard deviation of the elements in vector $D(a)$. That is, the standard deviation of distances one can go by successively following each of the children of a.

Similarly, the *child vector* $C(a)$ [Botafogo et al. 1992] is defined as follows:

$$C(a) = \begin{cases} \left\{1 + \sum C(a_1), 1 + \sum C(a_2), \ldots, 1 + \sum C(a_n)\right\}, \\ \{0\} \quad \text{if } a \text{ has no child } (n = 0), \end{cases}$$

where $\sum C(a_i)$ is the sum of all elements in vector $C(a_i)$. The child vector is represented inside braces { }, and indicates the size (number of elements) of the subtrees rooted at $a_1, a_2, \ldots, a_n$. The *absolute child*

imbalance for node a is the standard deviation of the elements in vector $C(a)$. That is, it is the standard deviation of the number of nodes in the subtrees rooted at the children of a.

2.4. Random Graph Models

The theory of random networks is concerned with the structure and evolution of large, intricate networks depicting the elements of complex systems and the interactions between them. For example, living systems form huge genetic networks whose vertices are proteins and edges represent the chemical interactions between them. Similarly, a large network is formed by the nervous system whose vertices are nerve cells or neurons connected by axons. In social science, similar networks can be perceived between individuals and organizations. We describe random networks here in the context of another obvious instance of a large and complex network—the WWW.

The earliest model of the topology of these networks, known as the *random graph model*, due to Erdos and Renyi, is described in Barabasi et al. [1999]. Suppose we have a fixed sized graph of N vertices and each pair of vertices is connected with a probability p. The probability $P(k)$ that a vertex has k edges is assumed to follow the Poisson distribution such that $P(k) = e_x^{(-\lambda)}\lambda^k/k!$ where the mean λ is defined as

$$\lambda = N\binom{N-1}{k}p^k(1-p)^{N-k-1}.$$

Several random networks such as the WWW exhibit what is known as the small-world phenomenon whereby the average distance between any pair of nodes is usually a small number. Measurements by Albert et al. [1999] show that, on average, two randomly chosen documents on the Web are a mere 19 clicks away. The *small-world model* accounts for this observation by viewing the N vertices as a one-dimensional lattice where each vertex is connected to its two nearest and next-nearest neighbors. Each edge is reconnected to a vertex chosen at random with probability p. The long range connections generated by this process decrease the distances between vertices, leading to the small-world phenomenon. Both the models predict that the probability distribution of vertex connectivity has an exponential cutoff.

Barabasi and Albert [1999] reported from their study of the topology of several large networks that, irrespective of the type of network, the connectivity distribution (or the probability $P(k)$ that a vertex in the network interacts with k other vertices) decays as a power law, that is, $P(k) \sim k^{-\gamma}$, where γ is a small constant usually between 2.1 and 4, depending on the network in question. This observation has been confirmed for the WWW by measurements of Web-page degree distributions [Kleinberg et al. 1999; Kumar et al. 1999]. The power law distribution suggests that the connectivity of large random networks is free of scale, an implication inconsistent with the traditional random network models outlined above.

The random network models outlined earlier do not incorporate two generic aspects of real networks. First, they assume that the number of vertices N remains fixed over time. In contrast, most real networks are constantly changing due to additions and deletions of nodes. Typically, the number of vertices N increases throughout the lifetime of the network. The number of pages in the WWW in particular is reportedly growing exponentially over time. Second, the random network models assume that the probability of an edge existing between two vertices is uniformally distributed. However, most real networks exhibit what is known as *preferential connectivity*, that is, newly added vertices are more likely to establish links with vertices having higher connectivity. In the WWW context, this manifests in the propensity of Web authors to include links to highly connected documents in their web pages. These ingredients form the basis of the *scale-free model* proposed by Barabasi and Albert [Albert and Barabasi 2000; Barabasi

et al. 1999, 2000]:

—To incorporate the growth of the network, starting with a small number m_0 of vertices, we add, at each time step, a new vertex with $m(< m_0)$ edges that link to m vertices already present in the network.

—To incorporate preferential attachment, we assume that the probability Π that a new vertex will be connected to vertex i is proportional to the relative connectivity of that vertex, that is, $\Pi(k_i) = k_i / \sum_j k_j$.

After t time steps, the above assumptions lead to a network with $t + m_0$ vertices and mt new edges. It can be shown that the network evolves to a scale-invariant state with the probability that a vertex has k edges following a power law with an exponent $\gamma \approx 3$, thereby reproducing the observed behavior of the WWW and other random networks. Note that the value of γ is dependent on the exact form of the growth and preferential attachment functions defined above and different values would be obtained if for instance the linear preferential attachment function were to be replaced by an exponential function.

3. WEB PAGE SIGNIFICANCE

Perhaps the most well-known Web metrics are *significance* metrics. The significance of a web page can be viewed from two perspectives—its *relevance* to a specific information need, such as a user query, and its absolute *quality* irrespective of particular user requirements. Relevance metrics relate to the similarity of Web pages with driving queries using a variety of models for performing the comparison. Quality metrics typically use link information to distinguish frequently referred pages from less visible ones. However, as we shall see, the quality metrics discussed here are more sophisticated than simple in-degree counts. The most obvious use of significance metrics is in Web search and retrieval where the most relevant and high quality set of pages must be selected from a vast index in response to a user query. The introduction of quality metrics has been a recent development for public search engines, most of which relied earlier on purely textual comparisons of keyword queries with indexed pages for assigning relevance scores. Engines such as Google [Brin and Page 1998] use a combination of relevance and quality metrics in ranking the responses to user queries.

3.1. Relevance

Information retrieval techniques have been adapted to the Web for determining relevance of web pages to keyword queries. We present four algorithms for relevance ranking as discussed by Yuwono and Lee [1996]; namely, *Boolean spread activation*, *most-cited*, TFxIDF, and *vector spread activation*. The first two rely on hyperlink structure without considering *term frequencies* (to be explained later.) The latter two are based on the *vector space model*, which represents documents and queries as vectors for calculating their similarity. Strictly speaking, relevance is a subjective notion as described in Yuwono and Lee [1996]:

> "... a WWW page is considered relevant to a query if, by accessing the page, the user can find a resource (URL) containing information pertinent to the query, or the page itself is such a resource."

As such, the relevance score metrics detailed below are means to identify web pages that are potentially useful in locating the information sought by the user. We first introduce the notation to be used in defining the relevance metrics.

M	Number of query words
Q_j	The jth query term, for $1 \leq j \leq M$
N	Number of WWW pages in index
P_i	The ith page or its ID
$R_{i,q}$	Relevance score of P_i with respect to query q
$Li_{i,k}$	Occurrence of an incoming link from P_k to P_i
$Lo_{i,k}$	Occurrence of an outgoing link from P_i to P_k
$X_{i,j}$	Occurrence of Q_j in P_i

3.1.1. Boolean Spread Activation. In the *Boolean model*, the relevance score is simply the number of query terms that

appear in the document. Since only conjunctive queries can be ranked using this model, disjunctions and negations have to be transformed into conjunctions. The Boolean spread activation extends this model by propagating the occurrence of a query word in a document to its neighboring documents.[3] Thus,

$$R_{i,q} = \sum_{j=1}^{M} I_{i,j},$$

where

$$I_{i,j} = \begin{cases} c_1 & \text{if } X_{i,j} = 1 \\ c_2 & \text{if there exists } k \text{ such that } X_{k,j} = 1 \text{ and } Li_{i,k} + Lo_{i,k} > 0 \\ 0 & \text{otherwise.} \end{cases}$$

The constant c_2 ($c_2 < c_1$) determines the (indirect) contribution from neighboring documents containing a query term. We may further enhance the Boolean spread activation model of Yuwono and Lee [1996] through two (alternative) recursive definitions of the contribution of each query term $I_{i,j}$ as follows:

(1) $I_{i,j} = \begin{cases} 1 & \text{if } X_{i,j} = 1 \\ cI_{k,j} & \text{if there exists } k \text{ such that } 0 < c < 1, X_{k,j} = 1 \text{ and } Li_{k,j} + Lo_{k,j} > 0; \end{cases}$

(2) $I_{i,j} = X_{i,j} + cI_{k,j}; 0 < c < 1$ and $Li_{k,j} + Lo_{k,j} > 0$.

Note that the above definition is recursive as the rank of a page is a function of whether the search term appears in it (term $X_{i,j}$) or in a page that it is connected to ($cI_{k,j}$) through in- or outlinks. The further the page that contains the search term is from the given page, the lower its contribution to the score (due to the positive coefficient $c < 1$). Although the definition recursively states the rank based on contribution of neighboring pages (which in turn use their neighbors), the computation can indeed be done iteratively (because it is a case of tail recursion). The choice of implementation is not a subject of discussion here.

[3] Assuming that documents linked to one another have some semantic relationships.

3.1.2. Most-Cited. Each page is assigned a score, which is the sum of the number of query words contained in other pages having a hyperlink referring to the page (citing). This algorithm assigns higher scores to referenced documents rather than referencing documents.

$$R_{i,q} = \sum_{k=1,k\neq i}^{N} \left(Li_{i,k} \sum_{j=1}^{M} X_{k,j} \right).$$

The most-cited relevance metric may be combined with the recursive definition of Boolean spread activation to overcome the above problem as follows:

$$I_{i,j} = X_{i,j} + c(Li_{k,j} + Lo_{k,j})I_{k,j};$$
$$0 < c < 1.$$

We note two benefits of the $(Li_{k,j} + Lo_{k,j})$ coefficient. First, the new metric is unbiased with regard to citing and cited pages. Second, the contribution from neighboring pages is scaled by the degree of connectivity.

3.1.3. TFxIDF. Based on the vector space model, the relevance score of a document is the sum of weights of the query terms that appear in the document, normalized by the Euclidean vector length of the document. The weight of a term is a function of the word's occurrence frequency (also called the *term frequency* (*TF*)) in the document and the number of documents containing the word in collection (the *inversedocument*

Table III. The Average Precision for Each Search Algorithm

	Boolean spread activation	*most-cited*	*TFxIDF*	*vector spread activation*
Average precision	0.63	0.58	0.75	0.76

frequency (*IDF*)).

$$R_{i,q} = \frac{\sum_{Q_j}(0.5 + 0.5(TF_{i,j}/TF_{i,max}))\, IDF_j}{\sqrt{\sum_{j \in P_i}(0.5 + 0.5(TF_{i,j}/TF_{i,max}))^2\, (IDF_j)^2}},$$

where

$TF_{i,j}$	Term frequency of Q_j in P_i
$TF_{i,max}$	Maximum term frequency of a keyword in P_i
IDF_j	$\log(\frac{N}{\sum_{i=1}^{N} X_{i,j}})$.

The weighing function (product of *TF* and *IDF* as in Lee et al. [1997]) gives higher weights to terms that occur frequently in a small set of documents. A less-expensive evaluation of the relevance score leaves out the denominator in the above (i.e., the normalization factor). Performance of several approximations of relevance score by the vector-space model is considered in Lee et al. [1997].

3.1.4. Vector Spread Activation. The vector space model is often criticized for not taking into account hyperlink information as was done in web page quality models [Brin and Page 1998; Kleinberg 1998; Marchiori 1997]. The vector spread activation method incorporates score propagation as done in Boolean spread activation. Each Web page is assigned a relevance score (according to the TFxIDF model) and the score of a page is propagated to those it references. That is, given the score of P_i as $S_{i,q}$ and the link weight $0 < \alpha < 1$,

$$R_{i,q} = S_{i,q} + \sum_{j=1, j \neq i}^{N} \alpha Li_{i,j} S_{j,q}.$$

However, experiments [Yuwono and Lee 1996] show that the vector spread activation model performs only marginally better than **TFxIDF** in terms of retrieval effectiveness measured by *precision* and *recall* (to be introduced later). Table III summarizes the average precision for each search algorithm for 56 test queries as highlighted in Yuwono and Lee [1996].

Application of the above query relevance models in search services can be enhanced through relevance feedback [Lee et al. 1997; Yuwono et al. 1995]. If the user is able to identify some of the references as relevant, then certain terms from these documents can be used to reformulate the original query into a new one that may capture some concepts not explicitly specified earlier.

3.2. Quality

Recent work in Web search has demonstrated that the quality of a Web page is dependent on the hyperlink structure in which it is embedded. Link structure analysis is based on the notion that a link from a page p to page q can be viewed as an endorsement of q by p, and as some form of positive judgment by p of q's content. Two important types of techniques in link-structure analysis are *co-citation* based schemes and *random-walk*-based schemes. The main idea behind co-citation based schemes is the notion that, when two pages p_1 and p_2 both point to some page q, it is reasonable to assume that p_1 and p_2 share a mutual topic of interest. Likewise, when p links to both q_1 and q_2, it is probable that q_1 and q_2 share some mutual topic. On the other hand, random-walk-based schemes model the Web (or part of it) as a graph where pages are nodes and links are edges, and apply some

Table IV. Evaluation of Search Engines vs. their Hyper Versions

	Excite	*HotBot*	*Lycos*	*WebCrawler*	*OpenText*	*Average*
Evaluation increment	+5.1	+15.1	+16.4	+14.3	+13.7	+12.9
Std Deviation	2.2	4.1	3.6	1.6	3	2.9

random walk model to the graph. Pages are then ranked by the probability of visiting them in the modeled random walk.

In this section, we discuss some of the metrics for link structure analysis. Each of these metrics is recursively defined for a Web page in terms of the measures of its neighboring pages and the degree of its hyperlink association with them. Quality metrics can be used in conjunction with relevance metrics to rank results of keyword searches. In addition, due to their independence from specific query contexts, they may be used generically for a number of purposes. We mention these applications in the context of individual quality metrics. Also, page quality measures do not rely on page contents which make them convenient to ascertain and at the same time sinister "spamdexing" schemes[4] becomes relatively more difficult to implement.

3.2.1. Hyper-information Content. According to Marchiori [1997], the *overall information* of a web object is not composed only by its static *textual information*, but also *hyper* information, which is the measure of the potential information of a Web object with respect to the Web space. Roughly, it measures how much information one can obtain using that page with a browser, and navigating starting from it. Suppose the functions $I(p)$, $T(p)$ and $H(p)$ denote the overall information, textual information and hyper information, respectively, of p that map Web pages to nonnegative real numbers with an upper bound of 1. Then,

$$I(p) = T(p) + H(p).$$

Then it can be shown that given a page p that points to q, and a suitable fading factor $f(0 < f < 1)$, we have

$$I(p) = T(p) + fT(q).$$

It can be shown that the contribution of a page q_n at a distance of n clicks from p to p's information $I(p)$ is $f^n T(q_n)$. To model the case when p contains several hyperlinks without violating the bounded property of $I(p)$, the *sequential selection* of links $b_1, b_2, \ldots, b_n$ embedded in p is interpreted as follows:

$$I(p) = T(p) + fT(b_1) + \cdots + f^n T(b_n).$$

Therefore, the sequential selection of links contained in p is interpreted in the same way as following consecutive links starting at p.

In order to validate the above notion of hyper information, Marchiori [1997] implemented the hyper information as post-processor of the search engines available during the time of his work (Excite, HotBot, Lycos, WebCrawler and OpenText). The post-processor remotely query the search engines, extract the corresponding scores ($T(p)$ function), and calculate the hyper information and therefore the overall information by fixing the depth and fading factors in advance. Table IV [Marchiori 1997] shows the evaluation increment for each search engine with respect to its hyper version and the corresponding standard deviation. As can be seen, the small standard deviations are empirical evidence of the improvement of the quality of information provided by the hyper search engines over their nonhyper versions.

3.2.2. Impact Factor. The impact factor, which originated from the field of *bibliometrics*, produces a quantitative estimate of the significance of a scientific journal. Given the inherent similarity of web sites to journals arising from the analogy

[4] The judicious use of strategic keywords that makes pages highly visible to search engine users irrespective of the relevance of their contents.

between page hyperlinks and paper citations, the impact factor can also be used to measure the significance of web sites. The *impact factor* [Egghe and Rousseau 1990] of a journal is the ratio of all citations to a journal to the total number of source items (that contain the references) published over a given period of time. The number of citations to a journal or its in-degree is limited in depicting its standing. As pointed out in [Egghe and Rousseau 1990], it does not contain any correction for the average length of individual papers. Second, citations from all journals are regarded as equally important. A more sophisticated citation-based impact factor than normalized in-degree count as proposed by Pinski and Narin is discussed in Egghe and Rousseau [1990] and Kleinberg [1998]. We follow the more intuitive description in Kleinberg [1998].

The impact of a journal j is measured by its *influence weight* w_j. Modeling the collection of journals as a graph, where the nodes denoting journals are labeled by their influence weights and directed edges by the connection strength between two nodes, the connection strength S_{ij} on the edge $\langle i, j \rangle$ is defined as the fraction of citations from journal i to journal j. Following the definitions above, the influence weight of journal j is the sum of influence weights of its citing journals scaled by their respective connection strengths with j. That is,

$$w_j = \sum_i w_i S_{ij}.$$

The nonzero, nonnegative solution w to the above system of equations ($w = S^T w$) is given by the principal eigenvector[5] of S^T.

3.2.3. PageRank. The PageRank measure [Brin and Page 1998] extends other citation based ranking measures, which merely count the citations to a page. Intuitively, a page has a high PageRank if there are many pages that point to it or if there are some pages with high PageRank that point to it. Let N be the set of pages that point to a page p and $C(p)$ be the number of links going out of a page p. Then, given a damping factor d, $0 \leq d \leq 1$, the PageRank of p is defined as follows:

$$R(p) = (1 - d) + d \sum_{q \in N} \frac{R(q)}{C(q)}.$$

The PageRank may also be considered as the probability that a *random surfer* [Brin and Page 1998] visits the page. A random surfer who is given a Web page at random, keeps clicking on links, without hitting the "back" button but eventually gets bored and starts from another random page. The probability that the random surfer visits a page is its PageRank. The damping factor d in $R(p)$ is the probability at each page the random surfer will get bored and request for another random page. This ranking is used as one component of the Google search engine [Brin and Page 1998] to help determine how to order the pages returned by a Web search query. The score of a page with respect to a query in Google is obtained by combining the position, font and capitalization information stored in *hitlists* (the IR score) with the PageRank measure. User feedback is used to evaluate search results and adjust the ranking functions. Cho et al. [1998] describe the use of PageRank for ordering pages during a crawl so that the more important pages are visited first. It has also been used for evaluating the quality of search engine indexes using random walks [Henzinger et al. 1999].

3.2.4. Mutual Reinforcement Approach. A method that treats hyperlinks as conferrals of authority on pages for locating relevant, authoritative WWW pages for a broad topic query is introduced by Kleinberg in [1998]. He suggested that Web-page importance should depend on the search query being performed. This model is based on a mutually reinforcing

[5] Let M be a $n \times n$ matrix. An *eigenvalue* of M is a number λ with the property that, for some vector ω, we have $M\omega = \lambda\omega$. When the assumption that $|\lambda_1(M)| > |\lambda_2(M)|$ holds, $\omega_1(M)$ is referred to as the *principal eigenvector* and all other $\omega_i(M)$ as *nonprincipal eigenvectors*.

relationship between *authorities*—pages that contain a lot of information about a topic, and *hubs*—pages that link to many related authorities. That is, each page should have a separate *authority* rating based on the links going to the page and *hub* rating based on the links going from the page. Kleinberg proposed first using a text-based Web search engine to get a Root Set consisting of a short list of Web pages relevant to a given query. Second, the Root Set is augmented by pages that link to pages in the Root Set, and also pages that are linked from pages in the Root Set, to obtain a larger Base Set of Web pages. If N is the number of pages in the final Base Set, then the data of Kleinberg's algorithm consists of an $N \times N$ adjacency matrix A, where $A_{ij} = 1$ if there are one or more hypertext links from page i to page j; otherwise, $A_{ij} = 0$.

Formally, given a focused subgraph that contains a relatively small number of pages relevant to a broad topic,[6] the following rule is used to iteratively update authority and hub weights (denoted x_p and y_p, respectively, and initialized to 1) of a page p:

$$x_p = \sum_{q:q\to p} y_q \quad \text{and dually}$$

$$y_p = \sum_{q:p\to q} x_q$$

The weights are normalized after each iteration to prevent them from overflowing. At the end of an arbitrarily large number of iterations, the authority and hub weights converge to fixed values. Pages with weights above a certain threshold can then be declared as authorities and hubs respectively. If we represent the focused subgraph as an adjacency matrix A where $A_{ij} = 1$ if there exists a link from page i to page j, and $A_{ij} = 0$ otherwise, it has been shown that the authority and hub vectors ($x = \{x_p\}$ and $y = \{y_p\}$) converge to the principal eigenvectors of $A^T A$ and AA^T respectively [Kleinberg 1998].

[6] The focused subgraph is constructed from a *Root Set* obtained from a search engine query.

Authority and hub weights can be used to enhance Web search by identifying a small set of high-quality pages on a broad topic [Chakrabarti et al. 1998a, 1998b]. Pages related to a given page p can be found by finding the top authorities and hubs among pages in the vicinity to p [Dean and Henzinger 1999]. The same algorithm has also been used for finding densely linked communities of hubs and authorities [Gibson et al. 1998].

One of the limitations of Kleinberg's [1998] *mutual reinforcement principle* is that it is susceptible to the *Tightly Knit Communities* (*TKC*) effect. The TKC effect occurs when a community achieves high scores in link-analysis algorithms even as sites in the TKC are not authoritative on the topic, or pertain to just one aspect of the topic. A striking example of this phenomenon is provided by Cohn and Chang [2000]. They use Kleinberg's Algorithm with the search term "jaguar," and converge to a collection of sites about the city of Cincinnati! They found out that the cause of this is a large number of online newspaper articles in the Cincinnati Enquirer which discuss the Jacksonville Jaguars football team, and all link to the same Cincinnati Enquirer service pages.

3.2.5. Rafiei and Mendelzon's Approach. Generalizations of both PageRank and authorities/hubs models for determining the topics on which a page has a reputation are considered by Rafiei and Mendelzon [2000]. In the one-level influence propagation model of PageRank, a surfer performing a random walk may jump to a page chosen uniformally at random with probability d *or* follow an outgoing link from the current page. Rafiei and Mendelzon [2000] introduce into this model, topic specific surfing and parameterize the step of the walk at which the rank is calculated. Given that N_t denotes the number of pages that address topic t, the probability that a page p will be visited in a random jump during the walk is d/N_t if p contains t and zero otherwise. The probability that the surfer visits p after n steps, following

a link from page q at step $n-1$ is $((1-d)/O(q))R^{n-1}(q,t)$, where $O(q)$ is the number of outgoing links in q and $R^{n-1}(q,t)$ denotes the probability of visiting q for topic t at step $n-1$. The stochastic matrix containing pairwise transition probabilities according to the above model can be shown to be aperiodic and irreducible, thereby converging to stationary state probabilities when $n \to \infty$. In the two-level influence propagation model of authorities and hubs [Kleinberg 1998], outgoing links can be followed *directly* from the current page p, or *indirectly* through a random page q that has a link to p.

3.2.6. SALSA. Lempel and Morgan [2000] propose the Stochastic Approach for Link Structure Analysis (SALSA). This approach is based upon the theory of Markov Chains, and relies on the stochastic properties of random walks[7] performed on a collection of sites. Like Kleinberg's algorithm, SALSA starts with a similarly constructed Base Set. It then performs a random walk by alternately (a) going uniformly to one of the pages that links to the current page, and (b) going uniformly to one of the pages linked to by the current page. The authority weights are defined to be the stationary distribution of the two-step chain doing first step (a) and then (b), while the hub weights are defined to be the stationary distribution of the two-step chain doing first step (b) and then (a).

Formally, let $B(i) = \{k : k \to i\}$ denote the set of all nodes that point to i, and let $F(i) = \{k : i \to k\}$ denote the set of all nodes that we can reach from i by following a forward link. It can be shown that the Markov Chain for the authorities has transition probabilities

$$P_a(i,j) = \sum_{k:k \in B(i) \cap B(j)} \frac{1}{|B(i)|} \frac{1}{|F(k)|}.$$

[7] According to [Rafiei and Mendelzon 2000], a *random walk* on a set of states $S = \{s_1, s_2, \ldots, s_n\}$, corresponds to a sequence of states, one for each step of the walk. At each step, the walk switches to a new state or remains in the current state. A random walk is *Markovian* if the transition at each step is independent of the previous steps and only depends on the current state.

Assume for the time being the Markov Chain is irreducible, that is, the underlying graph structure consists of a single connected component. The authors prove that the stationary distribution $a = (a_1, a_2, \ldots, a_N)$ of the Markov Chain satisfies $a_i = |B(i)|/|B|$, where $B = \bigcup_i B(i)$ is the set of all backward links. Similarly, the Markov Chain for the hubs has transition probabilities

$$P_h(i,j) = \sum_{k:k \in F(i) \cap F(j)} \frac{1}{|F(i)|} \frac{1}{|B(k)|}.$$

Lempel and Moran [2000] proved that the stationary distribution $= (h_1, h_2, \ldots, h_N)$ of the Markov Chain satisfies $h_i = |F(i)|/|F|$, where $F = \bigcup_i F(i)$ is the set of all forward links.

If the underlying graph of the Base Set consists of more than one component, then the SALSA algorithm selects a starting point uniformly at random and performs a random walk within the connected component that contains the node.

Observe that SALSA does not have the same *mutually reinforcing structure* that Kleinberg's algorithm does [Borodin et al. 2001]. Since $a_i = |B(i)|/|B|$, the relative authority of site i within a connected component is determined local links, not from the structure of the component. Also, in the special case of a single component, SALSA can be viewed as a one-step truncated version of Kleinberg's algorithm [Borodin et al. 2001]. Furthermore, Kleinberg ranks the authorities based on the structure of the entire graph, and tends to favor the authorities of tightly knit communities. The SALSA ranks the authorities based on the their popularity in the immediate neighborhood, and favors various authorities from different communities. Specifically, in SALSA, the TKC effect is overcome through random walks on a bipartite Web graph for identifying authorities and hubs. It has been shown that the resulting Markov chains are ergodic and high entries in the stationary distributions represent sites most frequently visited in the random walk. If the Web graph is weighted, the authority

and hub vectors can be shown to have stationary distributions with scores proportional to the sum of weights on incoming and outgoing edges, respectively. This result suggests a simpler calculation of authority/hub weights than through the mutual reinforcement approach.

3.2.7. Approach of Borodin et al. Borodin et al. [2001] proposed a set of algorithms for hypertext link analysis. We highlight some of these algorithms here. The authors proposed a series of algorithm that are based on minor modification of Kleinberg's algorithm to eliminate the previously mentioned errant behavior of Kleinberg's algorithm. They proposed an algorithm called *Hub-Averaging-Kleinberg Algorithm*, which is a hybrid of the Kleinberg and SALSA algorithms as it alternated between one step of each algorithm. It does the authority rating updates just like Kleinberg (giving each authority a rating equal to the sum of the hub ratings of all the pages that link to it). However, it does the hub rating updates by giving each hub a rating equal to the average of the authority ratings of all the pages that it links to. Consequently, a hub is better if it links to only *good* authorities, rather than linking to both good and bad authorities. Note that it shares the following behavior characteristics with the Kleinberg algorithm: if we consider a full bipartite graph, then the weights of the authorities increase exponentially fast for Hub-Averaging (the rate of increase is the square root of that of the Kleinberg's algorithm). However, if one of the hubs point to a node outside the component, then the weights of the component drop. This prevents the Hub-Averaging algorithm from completely following the drifting behavior of the Kleinberg's algorithm [Borodin et al. 2001]. Hub-Averaging and SALSA also share a common characteristic as the Hub-Averaging algorithm tends to favor nodes with high in-degree. Namely, if we consider an isolated component of one authority with high in-degree, the authority weight of this node will increase exponentially faster [Borodin et al. 2001].

The authors also proposed two different algorithms called *Hub-Threshold* and *Authority-Threshold* that modifies the "threshold" of Kleinberg's algorithm. The *Hub-Threshold algorithm* is based on the notion that a site should not be considered a good authority simply because many hubs with very poor hub weights point to it. When computing the authority weight of ith page, the Hub-Threshold algorithm does not take into consideration all hubs that point to page i. It only considers those hubs whose hub weight is at least the average hub weight over all the hubs that point to page i, computed using the current hub weights for the nodes.

The *Authority-Threshold algorithm*, on the other hand, is based on the notion that a site should not be considered a good hub simply because it points to a number of "acceptable" authorities; rather, to be considered a good hub it must point to some of the best authorities. When computing the hub weight of the ith page, the algorithm counts those authorities which are among the top K authorities, based on the current authority values. The value of K is passed as a parameter to the algorithm.

Finally, the authors also proposed two algorithms based on Bayesian statistical approach, namely, *Bayesian Algorithm* and *Simplified Bayesian Algorithm*, as opposed to the more common algebraic/graph theoretic approach. The Simplified Bayesian Algorithm is basically simplification of the Bayesian model in the Bayesian algorithm. They experimentally verified that the *Simplified Bayesian Algorithm* is almost identical to the SALSA algorithm and have at least 80% overlap on all queries. This may be due to the fact that both the algorithms place great importance on the in-degree of a node while determining the authority weight of a node. On the other hand, the *Bayesian algorithm* appears to resemble both the Kleinberg and the SALSA behavior, leaning more towards the first. It has a higher intersection numbers with Kleinberg than with SALSA [Borodin et al. 2001].

3.2.8. PicASHOW. PicASHOW [Lempel and Soffer 2001] is a pictorial retrieval

system that searches for images on the Web using hyperlink-structure analysis. PicASHOW applies co-citation based approaches and PageRank influenced methods. It does not require any image analysis whatsoever and no creation of taxonomies for preclassification of the images on the Web. The justification for using co-citation based measures to images just as it does to Web pages is as follows: (1) Images that are co-contained in pages are likely to be related to the same topic. (2) Images that are contained in pages that are co-cited by a certain page are likely related to the same topic. Furthermore, in the spirit of PageRank, the authors assumed that images that are contained in authoritative pages on topic t are good candidates to be quality images on that topic.

The topical collection of images from the Web is formally defined as a quadruple $IC = (P, I, L, E)$, where P is a set of Web pages (many of which deal with a certain topic t), I is the set of images contained in P, $L \subseteq P \times P$ is the set of directed links that exist on the Web between the pages P, and $E \subseteq P \times I$ is the relation page p *containing* image i. A page p contain an image i if (a) when p is loaded in a Web browser, i is displayed, or (b) p points to i's image file (in some image file format such as .gif or .jpeg). Based on this definition of image collection, the steps for finding authoritative images given in a query are as follows:

(1) The first step is to assemble a large topical collection of images for a given query on topic t. This is based on the notion that by examining a large enough set of t-relevant pages, it is possible to identify high quality t-images. This is achieved by using Kleinberg's algorithm. That is, for a query q on topic t, Kleinberg's algorithm is used to assemble a q-induced collection of Web pages by submitting q first to traditional search engines, and adding pages that point to or pointed by pages in the resultant set. This is the page set P and the page-to-page link set L. The set of images I can be then defined by collecting the images which are contained in P.
(2) The next step focuses on identifying replicated images in the collection. This is based on the assumption that when a web-site creator encounters an image of his liking on a remote server, the usual course of action is to copy the image file to the local server, thus replicating the image. Note that this behavior is different from the corresponding behavior with respect to HTML pages. Most of the time, authors will not copy a remote page to the local servers, but rather provide links from their site to the remote page. In PicASHOW, Lempel and Soffer [2001] download only the first 1024 bytes of the image and apply a double hash function to these bytes, so that each image is represented by a signature consisting of 32 bytes. Then, two images with the same signature are considered identical.
(3) Next, *noninformative* images are filtered out from the collection. The authors use the following heuristics in order to reduce noninformative images in the collection: (a) Banners and logos are noninformative and tend to be wide and short. Hence, PicASHOW filters out images with an aspect ratio greater than some threshold value. (b) Images that store small files (less than 10 kilobytes) tend to be banners and filtered out. Even if these files are not banners, they are usually not quality topical images. (c) Images stored in file names containing "logo" or "banner" keyword. (d) Additionally, cliparts, buttons, spinning globes are filtered out based on aspect ratio and file size heuristics. Note that this approach does not eliminate all noninformative images.
(4) Finally, the images are ranked based on different schemes such as in-degree approach, PageRank-influenced approach and co-citation based scheme.

Similarly, PicASHOW also finds *image hubs*. Just as hubs were defined as pages that link to many authoritative pages, *image hubs* are pages that are linked to

many authoritative images. Specifically, pages that contain high-quality images are called *image containers*, while pages that point to good image containers are called *image hubs*. Thus, image hubs are removed from the authoritative images themselves, which are contained in the image containers. The co-citation based image retrieval schemes is used once again to find both image containers and image hubs.

4. USAGE CHARACTERIZATION

In this section, we consider the problem of modeling and predicting web page accesses. We begin by relating web page access modeling to prediction for efficient information retrieval on the WWW and consider a preliminary statistical approach that relies on the distribution of interaccess times. Two other approaches from different domains, namely, Markov processes and human memory, are also examined.

4.1. Access Prediction Problem

In the case of the WWW, access prediction has several advantages. It becomes a basis for improved quality of information access by prefetching documents that have a high likelihood of access. In the Web environment, prefetching can materialize as a client-side or sever-side function depending on whether prediction is performed using the access patterns of a particular user or those of the whole population. If access prediction is addressed within the framework of a more general modeling problem, there are added benefits from using the model in other contexts. Webmasters can apply such a model for studying trends in page accesses recorded in server logs to identify navigation patterns, improve site organization and analyze the effects of changes to their web sites.

While the issue is important for improving information quality on the Web, it poses several challenges. The scale, extent, heterogeneity, and dynamism of the Web even within a site make several approaches to predicting accesses possible. However, as we shall see, each has its own set of limitations and assumptions that must be kept in mind while applying it to a particular domain.

Let us first elucidate the general access prediction problem. Our basic premise is that page accesses should be predicted based on universally available information on past accesses such as server access logs. Given a document repository and history of past accesses, we would like to know which documents are more likely to be accessed within a certain interval and how frequently they are expected to be accessed. The information used for prediction, typically found in server logs comprises the time and URL of an HTTP request. The identity of the client is necessary only if access prediction is personalized for the client. From this information about past accesses, several predictor variables can be determined, for example, the frequency of accesses within a time interval and interaccess times.

4.2. Statistical Prediction

An obvious prediction is the time until the next expected access to a document, say A. The duration can be derived from a distribution of time intervals between successive accesses. This kind of statistical prediction relates a predictor variable or a set of predictor variables to access probability for a large sample assumed to be representative of the entire population. Future accesses to a document can then be predicted from the probability distribution using current measurements of its predictor variable(s). A variant of this approach is to use separate distributions for individual documents measured from past accesses.

Let us illustrate the above approach for temporal prediction using interaccess time. Suppose $f(t)$ is the access density function denoting the probability that a document is accessed at time t after its last access or its *interaccess time probability density*. Intuitively, the probability that a document is accessed depends on the time since its last access and duration into the future we are predicting. At any

arbitrary point in time, the probability that a document A is accessed at a time T from now is given by

$$\Pr\{A \text{ is accessed at } T\} = f(\delta + T),$$

where δ is the *age* or the time since the last access to the document. The function $f(t)$ has a cumulative distribution $F(t) = \sum_{t'=0}^{\infty} f(t')$, which denotes the probability that a document will be accessed *within* time t from now. Since $f(t)$ is a probability density, $F(\infty) = 1$, meaning that the document will certainly be accessed sometime in the future. If $f(t)$ is represented as a continuous distribution, the instantaneous probability when $\delta, T \rightarrow 0$ is zero, which makes short-term or immediate prediction difficult. To find the discrete density $f(t)$ from the access logs, we calculate the proportion of document accesses that occur t time units after the preceding access for t ranging from zero to infinity. This approach assumes that all documents have identical interaccess time distributions, that is, all accesses are treated the same, irrespective of the documents they involve and that the distributions are free from periodic changes in access patterns (such as weekends when interaccess times are longer.) The implication of the first assumption is that the prediction is not conditioned on frequency of past accesses since all documents in the observed repository are assumed equally likely to be accessed giving rise to identical frequency distributions.

Since frequency distributions are more likely to vary between documents than not, it is clear that the above assumptions make this analysis suitable only on a per-document basis. However, the approach still holds, notwithstanding that the distribution $F(t)$ is now specific to a particular document. To predict the probability of access within time T from now, for a particular document A, we may use A's distribution function $F_A(t)$ to obtain $F_A(\delta + T)$ where δ is the age at the current time. If the interaccess time distributions are similar but not identical, we could condition these distributions on the parameters and find distributions of these parameters across the documents.

Our use of a single predictor, the interaccess time, obtained from the age δ and prediction interval T does not imply that the technique is univariate. The use of multiple predictors, such as the frequency of accesses in a given previous interval can easily be accommodated into a multidimensional plot of access probability. The method becomes complicated when several dimensions are involved. To alleviate this, we may derive a combined metric from the predictor variables, transforming the problem back to univariate prediction. However, this requires empirical determination of correlation between predictors and subsequently a combination function.

Given the statistical principle, one might naturally be led to ask how the distribution $F(t)$ (or its variant $F_A(t)$) can be used for actionable prediction. Recall that $F(t)$ is a cumulative probability distribution. For a given document age, it tells us the probability that a document is accessed within a certain interval of time. If a single probability distribution is used, this probability is an indicator of overall document usage with respect to time interval. If we use individual distributions $F_A(t)$, it can be used to compare the relative usage of documents. The expected time to next access $\overline{T}$ is given by the mean of the distribution:

$$E[T] = \sum_{t=0}^{\infty} t \cdot f(t).$$

The expected time $\overline{T}$ before the next access to a document, if it is known for all documents, can be used as a criteria for populating server side caches.

The temporal approach discussed above bases prediction on interaccess times. Equally, we may use a frequency based alternative for predicting access. A *frequency distribution* denotes the probability of a certain number of accesses to a document or a sample of documents over a *fixed* time interval. Using an analogous method to that discussed earlier, we can

answer the following for prediction over the next time interval:

— What is the probability that exactly N documents will be accessed?
— What is the probability that N or more documents will be accessed?
— How many documents are expected to be accessed?

This approach has the same drawbacks as discussed previously—it does not account for periodic changes in access rates, rather it aggregates them into a single distribution and accesses to all documents are treated the same. Finally, both temporal and frequency prediction may be combined to ascertain probabilities of a certain number of accesses during a given time period in the future.

4.3. Markov Models

Markov models assume web page accesses to be "memoryless," that is, access statistics are independent of events more than one interval ago. We discuss two Markov models for predicting web page accesses, one of our own making and another due to Sarukkai [2000]. Prior to explaining these models per se, we introduce the concept of Markov processes.

4.3.1. Markov Processes. A *stochastic process* can be thought of as a sequence of states $\{S_t; t = 1, 2, \ldots\}$ where S_t represents the state of the process at discrete time t, with a certain *transition probability* p_{ij} between any two states i and j. A *Markov process* is a simple form of stochastic process where the probability of an outcome depends only on the immediately preceding outcome. Specifically, the probability of being in a certain state at time t depends entirely on the state of the process at time $t-1$. To find the state probabilities at a time t, it is therefore sufficient to know the state probabilities at $t-1$ and the one-step transition probabilities, p_{ij} defined as

$$p_{ij} = \Pr\{S_t = j \,|\, S_{t-1} = i\}.$$

The one-step transition probabilities represent a *transition matrix* $P = (p_{ij})$. In a Markov process where the transition probabilities do not change with time, that is, a *stationary Markov process*, the probability of a transition in n steps from state i to state j or $\Pr\{S_t = j \,|\, S_{t-n} = i\}$ denoted by $p_{ij}^{(n)}$ is given by P_{ij}^n. Intuitively, this probability may be calculated as the summation of the transition probabilities over all possible n-step paths between i and j in a graph that is equivalent to the transition matrix P, with the transition probability along any path being the product of all successive one-step transition probabilities. This is stated precisely by the well-known *Chapman–Kolmogorov identity* [Ross 1983]:

$$p_{ij}^{(m+n)} = \sum_h p_{ih}^{(m)} p_{hj}^{(n)}, \quad (m, n) = 1, 2, \ldots \tag{1}$$

$$= P_{ij}^{(m+n)}. \tag{2}$$

Hence, the matrix of n-step transition probabilities is simply the nth power of the one-step transition matrix P. If the initial state probabilities $\pi_i^0 = \Pr\{S_0 = i\}$ are known, we can use the n-step transition probabilities to find the state probabilities at time n as follows:

$$\pi_j^n = \sum_i \Pr\{S_n = j \,|\, S_0 = i\} \cdot \Pr\{S_0 = i\}$$

$$\pi_j^n = \sum_i p_{ij}^{(n)} \pi_i^0.$$

Representing the state probabilities at time n as a state distribution vector $\Pi^n = (\pi_i^n)$, the above equation can be written in vector notation as

$$\Pi^n = P^n \cdot \Pi^0. \tag{3}$$

The vector of initial state probabilities Π^0 is known as the *initial state distribution*. Typically, as $n \to \infty$, the initial state distribution becomes less relevant to the n-step transition probabilities. In fact, for large n the rows of P^n become identical to each other and to the steady state distribution Π^n. The steady state probability π_i^n as $n \to \infty$ can be interpreted as the fraction of time spent by the process in state i in the long run.

4.3.2. Markov Chain Prediction. Markov processes can be used to model individual browsing behavior. We propose a method [Dhyani 2001] for determining access probabilities of web pages within a site by modeling the browsing process as an ergodic Markov chain [Ross 1983]. A *Markov chain* is simply a sequence of state distribution vectors at successive time intervals, that is, $\langle \Pi^0, \Pi^1, \ldots, \Pi^n \rangle$. A Markov chain is *ergodic* if it is possible to go from every state to every other state in one or more transitions. The approach relies on *a-posteriori* transition probabilities that can readily be ascertained from logs of user accesses maintained by Web servers. Note that our method can be considered as a variant of the PageRank. The difference between PagerRank and our approach is that PageRank assumes a uniform distribution for following outlinks while we use observed frequencies. Also, our model does not perform random jumps to arbitrary Web pages.

Let us represent a web site as a collection of K states, each representing a page. The evolution of a user's browsing process may then be described by a stochastic process with a random variable X_n (at time $n = 1, 2, \ldots$) acquiring a value x_n from the state space. The process may be characterized as a Markov chain if the conditional probability of X_{n+1} depends only on the value of X_n and is independent of all previous values. Let us denote the nonnegative, normalized probability of transition from page i to page j as p_{ij}, so that, in terms of the random variable X_n,

$$p_{ij} = P(X_{n+1} = j \mid X_n = i).$$

The transition probabilities can be represented in a $K \times K$ transition matrix P. We can then derive the following:

— The ℓ-step transition probability $p_{ij}^{(\ell)}$ is the probability that a user navigates from page i to page j in ℓ steps.[8] This probability is defined generally by the Chapman–Kolmogorov identity of Eq. (1).

— In ergodic Markov chains, the amount of time spent in a state i is proportional to the steady state probability π_i. If user browsing patterns can be shown to be ergodic, then pages that occupy the largest share of browsing time can be identified. Let Π^0 denote initial state distribution vector of a Markov chain; the jth element of Π^0 is the probability that user is initially at page j.[9] The state distribution vector Π^n at time n can then be obtained from Eq. (3).

It can be shown that, for large n, each element of P^n approaches the steady-state value for its corresponding transition. We can thus say that irrespective of the initial state distribution, the transition probabilities of an ergodic Markov chain will converge to a stationary distribution, provided such a distribution exists. The mean recurrence time of a state j is simply the inverse of the state probability, that is, $1/\pi_j$.

We now briefly address the issue of obtaining transition probabilities p_{ij}. It can be shown that the transition probability can be expressed in terms of the a-posteriori probability τ_{ij}, which can be observed from browsing phenomena at the site as follows:

$$\tau_{ij} = \frac{\text{Number of accesses from } i \text{ to } j}{\text{Total number of accesses from } i \text{ to all its neighbors}}.$$

Accesses originating from outside the web site can be modeled by adding another node (say $*$) collectively representing outside pages outside. The transition probabilities p_{*i} denote the proportion of times users enter the web site at page i.

Observe that in relation to actual browsing, our model implies that the transition

[8] A more Web savvy term would be *click distance*.

[9] In the Markov chain model for Web browsing, each new access can be seen as an advancement in the discrete time variable.

to the next page is dependent only on the current page. This assumption correctly captures the situation whereby a user selects a link on the current page to go to the next in the browsing sequence. However, the model does not account for external random jumps by entering a fresh URL into the browser window or browser based navigation through the "Forward" and "Back" buttons. Note that pushing "Back" and "Forward" buttons imply that the next state in the surf process depends not only on the current page but also on the browsing history. Browser-based buttons can be factored into the model by treating them as additional links. However, we use a simplistic assumption because the a-priori probability of using browser buttons depends on the navigability of the web site and is difficult to determine by tracking sample user sessions (due to caching etc). Our model can be useful in the following applications:

—*Relative Web Page Popularity.* Since the transition probability matrix is known a-priori, the long term probabilities of visiting a page can be determined. This can be compared with other pages to find out which pages are more likely to be accessed than others.

—*Local Search Engine Ranking.* The relative popularity measure discussed above will make localized search engines more effective because it uses observed surfing behavior to rank pages rather than generalized in-out degrees. General purpose measures do not have this characteristic. For example, Pagerank assumes that out links are selected in accordance with a uniform distribution. However, hyperlinks often follow the 80/20 rule whereby a small number of links dominate the outgoing paths from a page. Long-term page probabilities that account for browsing behavior within the domain will give more accurate rankings of useful pages.

—*Smart Browsing.* Since ℓ-step transition probabilities are known, we can predict which pages are likely to be requested over the next ℓ steps. The browser can then initiate advanced access for them to improve performance. It can also serve suggestions to the user on which pages might be of interest based on their browsing history.

A variation of Markov chains applied by Sarukkai [2000] predicts user accesses based on the sequence of previously followed links. Consider a stochastic matrix P whose elements represent page transition probabilities and a sequence of vectors, one for each step in the link history of a user, denoted $I^1, I^2, \ldots I^{t-1}$. The ℓth element in vector I^k is set to 1 if the user visits page ℓ at time k, otherwise, it is set to 0. For appropriate values of constants $a_1, a_2, \ldots, a_k$, the state probability vector S^t for predicting the next link is determined as follows:

$$S_j^t = \sum_{k=1}^{t-1} a_k I^{t-k} P^k.$$

The next page to be accessed is predicted as the one with the highest state probability in the vector S^t. The same approach can be used to generate tours by successively predicting links of a path.

4.4. Human Memory Model

Psychology research has shown that the accuracy of recalling a particular item in memory depends on the number of times the item was seen or *frequency*, the time since last access or *recency* and the gap between previous accesses or *spacing*. In some ways, the human memory model is closer to Web-page access modeling. Recker and Pitkow [1996] have used the human memory model to predict document accesses in a multimedia repository based on the frequency and recency of past document accesses. We discuss their approach below.

4.4.1. Frequency Analysis. Frequency analysis examines the relationship between the frequency of document access during a particular time *window* and the probability of access during a subsequent marginal period called the *pane*. This relationship can be derived by observing

access log data showing the times of individual access to Web documents. Let us consider the number of documents accessed x times during a fixed interval, say one week. If we know the proportion of these documents that are also accessed during the marginal period, then we can calculate the probability of future access to a document given that it is accessed x times as follows. Suppose w_x denotes the number of documents accessed x times during the window period and p_x ($p_x < w_x$), the number, out of those accessed during the pane. Then, the probability that a document A is accessed during the pane given that it is accessed x times during the window is given by:

$$\Pr\{A \text{ is accessed in pane} \mid A \text{ is accessed } x \text{ times during window}\} = \frac{\Pr\{A \text{ is accessed in pane and } A \text{ is accessed } x \text{ times during window}\}}{\Pr\{A \text{ is accessed } x \text{ times during window}\}} = \frac{p_x}{w_x}.$$

The above probability can be aggregated over successive windows for the given observation interval to obtain the conditional distribution of *need probability*, or the probability that an item will be required right now, versus the frequency of access in the previous window. Recker and Pitkow found that the distribution of need probability p shows a familiar *Power Law* distribution to the human memory model, that is, as the frequency of document accesses x increases, the need probability increases according to the function:

$$p = ax^b,$$

where a and b are constants.

4.4.2. Recency Analysis. Recency analysis proceeds in a similar manner except that the probability of access during the pane is conditioned upon how recently the document was last accessed. Instead of creating bins of similar frequency values observed over successive windows, the documents are aggregated into bins of similar recency values. That is, if x documents were accessed an interval t ago (recency time) in the previous window, out of which x' are accessed during the next pane, then the need probability is calculated as follows:

$$\Pr\{A \text{ is accessed in pane} \mid A \text{ was accessed } t \text{ units ago}) = \frac{x'}{x}.$$

The need probability versus recency time distribution shows an *inverse* power law relationship in agreement with retention memory literature. In addition, recency analysis shows a much better fit on available access data than frequency.

Recker and Pitkow [1996] also study the relationship between the access distributions and hyperlink structure. They found that a high correlation exists between the recency of access, number of links and document centrality[10] in the graph structure. Specifically, documents that are less recently accessed, have fewer mean number of links per document, and lower measures of relative in- and out-degrees. Thus, recently accessed documents have higher overall interconnectivity. This analysis can be applied to optimal ordering for efficient retrieval. Documents with high-need probability can be positioned for faster access (caching strategies based on need probability) or more convenient (addition of appropriate hyperlinks).

There are several issues in Recker and Pitkow's [1996] approach that require careful consideration before such a model can be adopted for practical access prediction. First, the window and pane sizes are known to have a strong impact on the quality of prediction. No rule for an ideal setting of these parameters exists and significant trial and error may be

[10] See Section 2 for the definition of centrality and other hypertext graph measures.

needed to identify the right values. Second, the model does not account for continuous changes in Web repositories such as aging effects and the addition and deletion of documents. Finally, although the model considers multiple predictors, frequency and recency, the method of combining these parameters is not scalable. We must note that there is a difference between predictors and affecter variables. While frequency and recency of access are strong predictors, they cannot be established as factors affecting need probability due to the absence of a causal relationship. More fundamental document properties such visibility (connectivity) and relevance to popular topics are more likely to be factors that determine need probability. However, for prediction purpose, we believe that all the three factors considered above can be used.

4.5. Adaptive Web Sites

Usage characterization metrics attempt to model and measure user behavior from browsing patterns gathered typically from server log files. These metrics have facilitated *adaptive web sites*—sites that can automatically improve their organization and presentation by learning from visitor access patterns [Perkowitz and Etzioni 1997, 1998, 1999].

A metric that helps find collections of pages that are visited together in sessions is the *co-occurrence frequency* [Perkowitz and Etzioni 1999]. For a sequence of page accesses (obtainable from the server access log), the conditional probability $P(p|q)$ is the probability that a visitor will visit page p given that she has already visited q. The co-occurrence frequency between the pair $\langle p, q \rangle$ is then $\text{Min}(P(p|q), P(q|p))$. Connected components in a graph whose edges are labeled with the co-occurrence frequencies represent clusters of pages that are likely to be visited together. The *quality* of such clusters is defined as the probability that a user who has visited one page in a cluster also visits another page in the same cluster. In a related work, Yan et al. [1996] represent user sessions as vectors where the weight of the ith page is the degree of interest in it measured through actions such as the number of times the page is accessed or the time the user spends on it. Users can then be clustered on the basis of the similarity between their session vectors (measured as Euclidean distance or angle measure). As the user navigates a site, he is assigned to one or more categories based on the pages accessed so far. Pages in matching categories are included as suggestions on top of the HTML document returned to the user if they have not been accessed so far and are unlinked to the document. The same study finds that the distribution of time spent by a user on a page is roughly Zipfian. An analysis of user navigation patterns by Catledge and Pitkow [1995] reveals that the distribution of path lengths within a site is roughly negative linear with the relationship between path length p and its frequency f being $f = 0.24p$.

4.6. Activation Spread Technique

Pirolli et al. [1996] have recently used an activation spread technique to identify pages related to a set of "source" pages on the basis of link topology, textual similarity and usage paths. Conceptually, an activation is introduced at a starting set of web pages in the Web graph whose edges are weighted by the criteria mentioned above. To elaborate further, the degree of relevance of Web pages to one another is conceived as similarities, or strength of associations, among Web pages. These strength-of-association relations are represented using a composite of three graphs. Each graph structure contains nodes representing Web pages, and directed arcs among nodes are labeled with values representing strength of association among pages. These graphs represent the following:

— *The Hypertext Link Topology of a Web Locality.* This graph structure represents the hypertext link topology of a Web locality by using arcs labeled with unit strengths to connect one graph node to another when there exists a

hypertext link between the corresponding pages.

— *Inter-Page Text Similarity.* This type of graph structure represents the inter-page text content similarity by labeling arcs connecting nodes with the computed text similarities between corresponding Web pages.

— *Usage Paths.* The last type of graph structure represents the flow of users through the locality by labeling the arcs between two nodes with the number of users that go from one page to another.

Each of these graphs is represented by matrices in the spreading activation algorithm. That is, each row corresponds to a node representing a page, and similarly each column corresponds to a node representing a page. Conceptually, this activation flows through the graph, modulated by the arc strengths, the topmost active nodes represent the most relevant pages to the source pages. Effectively for a source page p, the asymptotic activity of page q in the network is proportional to $P(q|p)$, the probability that q will be accessed by a user given that she has visited p.

5. WEB PAGE SIMILARITY

Web page similarity metrics measure the extent of relatedness between two or more Web pages. Similarity functions have mainly been described in the context of Web page clustering schemes. Clustering is a natural way of semantically organizing information and abstracting important attributes of a collection of entities. Clustering has certain obvious advantages in improving information quality on the WWW. Clusters of Web pages can provide more complete information on a topic than individual pages, especially in an exploratory environment where users are not aware of several pages of interest. Clusters partition the information space such that it becomes possible to treat them as singular units without regarding the details of their contents. We must note, however, that the extent to which these advantages accrue depends on the quality and relevance of clusters. While this is contingent on user needs, intrinsic evaluations can often be made to judge cluster quality. In our presentation of clustering methods, we discuss these quality metrics where applicable.

We classify similarity metrics into *content-based*, *link-based* and *usage-based* metrics. Content-based similarity is measured by comparing the text of documents. Pages with similar content may be considered topically related and designated the same cluster. Link-based measures rely exclusively on the hyperlink structure of a Web graph to obtain related pages. Usage-based similarity is based on patterns of document access. The intent is to group pages or even users into meaningful clusters that can aid in better organization and accessibility of web sites.

5.1. Content-Based Similarity

Document resemblance measures in the Web context can use subsequences matching or word occurrence statistics. The first set of metrics using subsequence matching represents the document D as a set of fixed-size subsequences (or *shingles*) $S(D)$. The *resemblance* and *containment* [Broder et al. 1997] of documents are then defined in terms of the overlap between their shingle sets. That is, given a pair of documents A and B, the resemblance denoted ($r(A, B)$) and containment of A in B (denoted $c(A, B)$) are defined respectively as follows:

$$r(A, B) = \frac{|S(A) \cap S(B)|}{|S(A) \cup S(B)|}$$

$$c(A, B) = \frac{|S(A) \cap S(B)|}{|S(A)|}.$$

Clearly, both the measures vary between 0 and 1; if $A = B$, then a scalable technique to cluster documents on the Web using their resemblance is described in [Broder et al. 1997]. The algorithm locates clusters by finding connected components in a graph where edges denote the resemblance between document pairs.

Content-based similarity between Web pages can also be approached using the *vector space model* as in [Weiss et al. 1996]. Each document is represented as a term vector where the weight of term t_i in document d_k is calculated as follows:

$$w_{ki} = \frac{(0.5 + 0.5(TF_{k,i}/TF_{k,max}))w^{at}_{k,i}}{\sqrt{\sum_{t_i \in d_k}(0.5 + 0.5(TF_{k,i}/TF_{k,max}))^2 \left(w^{at}_{k,i}\right)^2}},$$

where

$TF_{k,i}$	Term frequency of t_i in d_k
$TF_{k,max}$	Maximum term frequency of a keyword in d_k
$w^{at}_{k,i}$	Contribution to weight from term attribute.

Note that this definition of term weights departs from Section 3.1 by excluding the inverse document frequency (IDF) and introducing a new factor w^{at} which is configurable for categories of terms. The term-based similarity, denoted S^t_{xy} between two documents d_x and d_y is the normalized dot product of their term vectors w_x and w_y respectively:

$$S^t_{xy} = \sum_i w_{xi} \cdot w_{yi}.$$

5.2. Link-Based Similarity

Link-based similarity metrics are derived from citation analysis. The notion of hyperlinks provides a mechanism for connection and traversal of information space as do citations in scholarly enterprise. Co-citation analysis has already been applied in information science to identify the core sets of articles, authors or journals in a particular field of study as well as for clustering works by topical relatedness. Although the meaning and significance of citations differs considerably in the two environments due to the unmediated nature of publishing on the Web, it is instructive to review metrics from citation analysis for possible adaptation. Application of co-citation analysis for topical clustering of WWW pages is described in Larson [1996] and Pitkow and Pirolli [1997]. We discuss here two types of citation based similarity measures namely *co-citation strength* and *bibliometric coupling strength* together with their refinements and applications to Web page clustering.

To formalize the definition of citation-based metrics, we first develop a matrix notation for representing a network of scientific publications. A bibliography may be represented as a *citation graph* where papers are denoted by nodes and references by links from citing document to the cited document. An equivalent matrix representation is called the *citation matrix* (C) where rows denote citing documents and columns cited documents. An element C_{ij} of the matrix has value one (or the number of references) if there exists a reference to paper j in paper i, zero otherwise.

Two papers are said to be *bibliographically coupled* [Egghe and Rousseau 1990] if they have one or more items of references in common. The *bibliographic coupling strength* of documents i and j is the number of references they have in common. In terms of the citation matrix, the coupling strength of i and j denoted $S^{B'}_{ij}$ is equal to the scalar product of their citation vectors C_i and C_j. That is,

$$\begin{aligned} S^{B'}_{ij} &= C_i \cdot C^T_j \\ &= (C \cdot C^T)_{ij}. \end{aligned}$$

In order to compare the coupling strength of different pairs of documents, the above metric may be unsensitized to the number of references through the following normalization (U is the unit vector of the same dimension as C):

$$S^B_{ij} = \frac{C_i \cdot C^T_j}{C_i \cdot U^T + C_j \cdot U^T}.$$

An alternate notion is that of *co-citation*. Two papers are co-cited if there exists a

third paper that refers to both of them. The *co-citation strength* is the frequency with which they are cited together. The relative co-citation strength is defined as follows:

$$S_{ij}^{C} = \frac{C_i^T \cdot (C_j^T)^T}{C_i^T \cdot U^T + C_j^T \cdot U^T}.$$

Bibliographic coupling (and co-citation) has been applied for clustering documents. Two criteria discussed by Egghe and Rousseau [1990] that can be applied to Web pages are as follows:

(A) A set of papers constitute a related group $G_A(P_0)$ if each member of the group has at least one coupling unit in common with a fixed paper P_0. That is, the coupling strength between any paper P and P_0 is greater than zero. Then, $G_A(P_0; n)$ denotes that subset of $G_A(P_0)$ whose members are coupled to P_0 with strength n.

(B) A number of papers constitute a related group G_B if each member of the group has at least one coupling unit to every other member of the group.

A measure of co-citation based cluster similarity function is the *Jaccard index*, defined as:

$$S_j(i, j) = \frac{coc(i, j)}{cit(i) + cit(j) - coc(i, j)},$$

where $coc(i, j)$ denotes the co-citation strength between documents i and j given by $C_i^T \cdot (C_j^T)^T$ and $cit(i) = C_i^T \cdot U^T$ and $cit(j) = C_j^T \cdot U^T$ are the total number of citations received by i and j, respectively. Note that Jaccard's index is very similar to the relative co-citation strength defined by us earlier. Another similarity function is given by *Salton's cosine equation* as follows:

$$S_s(i, j) = \frac{coc(i, j)}{\sqrt{cit(i) \cdot cit(j)}}$$

In most practical cases, it has been found that Salton's similarity strength value is twice as calculated by the Jaccard index.

Amongst other approaches to co-citation based clustering, Larson [1996] employs the multidimensional scaling technique to discover clusters from the raw co-citation matrix. Similarly, Pitkow and Pirolli [1997] apply a transitive cluster growing method once the pairwise co-citation strengths have been determined. In another application, Dean and Henzinger [1999] locate pages related to a given page by looking for siblings with the highest co-citation strength. Finally, a method described by Egghe and Rousseau [1990] finds connected components in *co-citation networks* whose edges are labeled by co-citation strengths of document pairs. The size and number of clusters (or cluster cohesiveness) can be controlled by removing edges with weights below a certain threshold co-citation frequency.

A generalization of citation-based similarity measures considers arbitrarily long citation paths rather than immediate neighbors. Weiss et al. [1996] introduce a weighted linear combination of three components as their hyperlink similarity function for clustering. For suitable values of weights W_d, W_a and W_s, the hyperlink similarity between two pages i and j is defined as:

$$S_{ij}^{l} = W_d S_{ij}^{d} + W_a S_{ij}^{a} + W_s S_{ij}^{s}.$$

We now discuss each of the similarity components S_{ij}^{d}, S_{ij}^{a} and S_{ij}^{s}. Let ℓ_{ij} denote the length of the shortest path from page i to j and l_{ij}^{k} that of the shortest path not traversing k.

—*Direct Paths.* A link between documents i and j establishes a semantic relation between the two documents. If these semantic relations are transitive, then a path between two nodes also implies a semantic relation. As the length of the shortest path between the two documents increases, the semantic relationship between the two documents tend to weaken. Hence, the direct path component relates the similarity between two pages i and j denoted S_{ij}^{s}

as inversely proportional to the shortest path lengths between them. That is,

$$S^s_{ij} = \frac{1}{2^{l_{ij}} + 2^{l_{ji}}}.$$

Observe that the denominator ensures that as shortest paths increase in length, the similarity between the documents decreases.

— *Common Ancestors.* This component generalizes co-citation by including successively weakening contributions from distant ancestors that are common to i and j. That is, the similarity between two documents is proportional to the number of ancestors that the two documents have in common. Let A denote the set of all common ancestors of i and j,

$$S^a_{ij} = \sum_{x \in A} \frac{1}{2^{l^j_{xi}} + 2^{l^i_{xj}}}.$$

Because direct paths have already been considered in similarity component S^s, only exclusive paths from the common ancestor to the nodes in question are involved in the S^a component. Observe that, as the shortest paths increase in length, the similarity decreases. Also, the more common ancestors, the higher the similarity.

— *Common Descendents.* This component generalizes bibliographic coupling in the same way as S^a. That is, the similarity between two documents is also proportional to the number of descendants that the two documents have in common. Let D denote the set of all common descendants of i and j,

$$S^d_{ij} = \sum_{x \in D} \frac{1}{2^{l^j_{ix}} + 2^{l^i_{jx}}}.$$

The computation normalizes S^d_{ij} to lie between 0 and 1 before it is included in S^l_{ij} in the same way as the normalization for S^a_{ij}.

5.3. Usage-Based Similarity

Information obtained from the interaction of users with material on the Web can be of immense use in improving the quality of online content. We now discuss some of the approaches proposed for relating Web documents based on user accesses. Web sites that automatically improve their organization and presentation by learning from access patterns are addressed by Perkowitz and Etzioni [1997, 1998, 1999]. Sites may be adaptive through *customization* (modifying pages in real time to suit individual users, e.g., goal recognition) or *optimization* (altering the site itself to make navigation easier for all, e.g., link promotion).

5.3.1. Clustering using Server Logs. The optimization approach is introduced by Perkowitz and Etzioni [1998, 1999] as an algorithm that generates a candidate index page containing clusters of web pages on the same topic. Their method assumes that a user visits conceptually related pages during an interaction with the site. The *index page synthesis problem* can be stated as follows: Given a web site and a visitor access log, create new index pages containing collections of links to related but currently unlinked pages. The PageGather cluster mining algorithm for generating the contents of the index page creates a small number of cohesive but possibly overlapping clusters through five steps:

(1) Process access logs into visits.
(2) Compute the co-occurrence frequencies between pages and create a similarity matrix. The matrix entry for a pair of pages p_i and p_j is defined as the co-occurrence frequency given by $\min(P(p_1|p_2), P(p_2|p_1))$. If two pages are already linked, the corresponding matrix cell is set to zero.
(3) Create a graph corresponding to the matrix and find the maximal cliques[11] or connected components[12] in the graph. Each clique or connected

[11] A *clique* is a subgraph in which every pair of nodes has an edge between them; a *maximum clique* is one that is not a subset of a larger clique.
[12] A *connected component* is a subgraph in which every pair of nodes has a path between them.

component then represents a set of pages that are likely to be visited together (and hence related, by the above assumption).

(4) Rank the clusters found and choose which to output. The clusters are sorted using the average co-occurrence frequency between all pairs of documents in a cluster.

(5) For each cluster, create a Web page consisting of links to the documents and present it to the Webmaster for evaluation.

Experiments [Perkowitz and Etzioni 1998, 1999] show that PageGather outperforms other clustering algorithms such as K-means, HAC, and a-priori frequency set calculation in efficiency as well as quality. Quality of clusters is measured by approximating the following: If the user visits a page in a cluster, what is the likelihood that he will visit more pages from the same cluster? Suppose $n(i)$ represents the number of pages in cluster i visited during a session, then the quality of cluster i is given by the probability $P\{n(i) \geq 2 | n(i) \geq 1\}$.

As we have seen, Perkowitz and Etzioni's approach relies on co-occurrence frequency to related pages. However, does frequent co-occurrence necessarily indicate semantic relationship? Certainly this assumption disregards a certain amount of arbitrariness that prevails in the often serendipitous browsing behavior of users. In exploratory browsing, pages accessed in the same session may not be related. For instance, high co-occurrence could also be due to site structure, such as coercive hyperlinks, rather than a persistent interest on the users behalf. The method discounts the nature of relationship between pages grouped together which might be a useful clue in obtaining more refined clusters.

Server logs can also help cluster users based on the similarity between the sets of pages they visit as described by Yan et al. [1996]. Clustering users is a natural way customization, whereby pages in a user's cluster that have not been explored yet can be suggested as navigational hints in the form of dynamically generated links. This type of dynamic hypertext configuration is performed as follows:

(1) *Preprocessing.* Suppose a web site has n HTML pages. Each user session[13] is represented using an n-dimensional *session vector* whose ith element is the *weight* or degree of interest assigned to the ith page. The weight can be measured through actions such as the number of times the page is accessed or the time the user spends on it.

(2) *Clustering.* User session vectors that are close to each other in the n-dimensional vector space, as determined by Euclidean or angular distance, are clustered together.

(3) *Dynamic Link Generation.* As the user navigates a site, she is assigned to one or more categories based on the pages accessed so far if their number exceeds a predefined threshold. Pages in matching categories are included as suggestions on top of the HTML document returned to the user if they have not been accessed so far and are unlinked to the document.

However, this approach towards customization assumes that

— Pages visited in each session are semantically related.

— Users will be interested in all pages accessed by other users in the same cluster (during dynamic hyperlink generation).

As for the optimization technique discussed earlier, these assumptions are not entirely valid. Similarity between session vectors does not necessarily imply a relationship between user interests. Both the techniques described above ignore the effect of proxies and browser caches on the accuracy of individual session tracking. Requests for cached pages are not logged by the server. Users behind a proxy or firewall cannot be distinguished and are assigned the same IP address, affecting

[13] A session is approximated as requests originating from the same host within 24 hours.

the granularity of clusters. Finally, estimates such as the number of accesses and time spent are not truly representative of a user's interest in a page.

6. WEB SEARCH AND RETRIEVAL

In Section 3, we introduced significance metrics for ranking responses of search engines to user queries. Here, we review metrics for evaluating and comparing the performance of Web search and retrieval services. We discuss them under two categories: *effectiveness* and *search engine comparison*.

6.1. Effectiveness

A large proportion of Web search engines employ information retrieval techniques such as the vector space model similarity for keyword based queries. Their retrieval effectiveness, or the quality of results returned in response to a query, is measured through two metrics, namely, *precision* and *recall*. Let N denote the number of responses of a search to a keyword query Q, of which N' are relevant to Q. If R relevant Web pages exist in all, then the effectiveness metrics are defined as follows [Hawking et al. 1999 Lee et al. 1997 Yuwono and Lee 1996]:

—*Precision.* The proportion of pages returned that are relevant. That is, Precision $= N'/N$.

—*Recall.* The proportion of relevant pages returned. That is, Recall $= N'/R$.

We assume here the definition of relevant pages stated in Section 3.1. Precision and recall usually exhibit a trade-off relationship that is depicted through Precision–Recall curves. Metrics that combine precision and recall are discussed in [Boyce et al. 1994]. One problem with measuring recall is that the total number of pages relevant to a query is difficult to ascertain given the size of the Web.

Recently, Brewington and Cybenko [2000] proposed a metric for the *currency* of a search engine. This metric can be used to answer questions about how fast a search engine must reindex the Web to remain "current." They introduce the concept of (α, β)-*currency* of a search engine with respect to a changing collection of Web pages. Informally, the search engine data for a given Web page is said to be β-current if the pages has not changed between the last time it was indexed and β time units ago. Formally, expected probability of a single page being β-current over all values of the observation time t_n is given as follows:

$$\Pr(\beta - current|\lambda, T, \beta) = \frac{\beta}{T} + \frac{1 - \exp(-\lambda(T - \beta))}{\lambda T},$$

where λ is the change rate of each Web page [Brewington and Cybenko 2000] and T is an associated distribution of reindexing times (a periodic reindexing system will have a single constant T_o). Observe that β is the grace period for allowing unobserved changes to a Web page and relaxes the temporal aspect of what it means to be current. The smaller β means more current our information about the page is.

A search engine for a collection of pages is then said to be (α, β)-current if a randomly chose page in the collection has a search engine entry that is β-current with probability at least α. It can be shown that the expected probability α is as follows:

$$\alpha = \int_0^\infty \left[\frac{\sigma}{\delta}\left(\frac{t}{\delta}\right)^{\sigma-1} \exp(-(t/\delta)^\sigma)\right] \times \left[\frac{\beta}{T_o} + \frac{1 - \exp(-(1/t)(T_o - \beta))}{(1/t)T_o}\right] dt,$$

where δ and σ are scale and shape parameters, respectively, in Weibull distributions [Montgomery and Runger 1994]. Note that the above integral can only be evaluated in closed form when the Weibull shape parameter σ is 1. Otherwise, numerical evaluation is required. Also, the integral gives an α for every pair of (T_o, β).

6.2. Search Engine Comparison

The comparison of public domain search services has been the subject of several

studies. One approach for comparing multiple search services describes the following metrics using a meta-search service such as MetaCrawler [Selberg and Etzioni 1995]:

— *Coverage.* Coverage measures the number of hits returned by each service on an average measured as the percentage of a pre-set maximum allowed hits per search engine. The uniqueness of coverage can be measured as the percentage of references not returned by any other engine. The absolute coverage of a public search engine or the fraction of the Web it indexes has been reported to be a maximum of one-third of the Web [Lawrence and Giles 1998, 1999]. The same study analyzes the overlap between pairs of engines to estimate a lower bound on the size of the indexable Web as 320 million pages. Note that the estimation of the size of the indexable Web is only valid till April 1998 [Lawrence and Giles 1998]. More recently, it has been estimated that the number of unique pages on the Internet is 2.1 billion and number of unique pages added per day is 7.3 million [Murray and Moore 2000]. These statistics are valid as of July 10, 2000.

— *Relevance.* Two relevance metrics are used. One metric is the proportion of hits from each engine that is followed by the user, or its precision. The other metric is the proportion of overall hits followed by the user per search engine (i.e., market share).

— *Performance.* Performance is measured by each service's average response time a query.

Other metrics to evaluate search engines indexes are *size*, *overlap* and *quality*. A standardized, statistical way of measuring relative search engine size and overlap using random queries (without privileged access) is described by Bharat and Broder [1998]. Suppose we have two procedures, one for sampling pages uniformly at random from the index of a particular search engine and another for checking whether a particular page is indexed by a particular search engine.

Based on approximations of these procedures through queries using public interfaces, the overlap and relative size of two search engines indexes E1 and E2 can be estimated as follows:

— *Overlap.* Fraction of URLs sampled from E1 found in E2. This approximates the following fraction:

$$\frac{|\mathrm{E1} \cap \mathrm{E2}|}{|\mathrm{E1}|}$$

— *Size Comparison.* The size ratio of E1 and E2 is given by:

$$\frac{\mathrm{size}(E1)}{\mathrm{size}(E2)} = \frac{\frac{|\mathrm{E1} \cap \mathrm{E2}|}{|\mathrm{E2}|}}{\frac{|\mathrm{E1} \cap \mathrm{E2}|}{|\mathrm{E1}|}} = \frac{\text{Fraction of URLs sampled from E2 found in E1}}{\text{Fraction of URLs sampled from E1 found in E2}}.$$

Sampling is performed by randomly selecting a URL from the pages returned for a query composed using keywords from a preconstructed lexicon. To test whether the URL is indexed by a particular search engine (i.e., checking), a *strong query*[14] (which uniquely identifies the page) is constructed. The presence of the URL is then tested in the set of pages returned. Note that the response to a strong query may

[14] A conjunctive query of k most significant keywords in the Web page. Significance is taken to be inversely proportional to frequency in the lexicon.

contain multiple URLs due to mirroring, multiple aliases, and so on.

A simple methodology for approximating index quality using random walks is proposed by Henzinger et al. [1999]. The definition of quality is based on the PageRank measure defined in the Google search engine [Brin and Page 1998]. Suppose each page on the Web is assigned a weight w scaled by the sum of all page weights. Then, the *quality* $w(S)$ and *average page quality* $a(S)$ of a search engine index S is defined as

$$w(S) = \sum_{p \in S} R(p)$$

$$a(S) = \frac{w(S)}{|S|},$$

where $R(p)$ is the PageRank[15] of a page p as defined in Brin and Page [1998]. This approach approximates the quality of a search engine index S by independently selecting pages $p_1, p_2, \ldots, p_n$ and testing whether each page is in S. Let $I[p_i \in S]$ be 1 if p_i is in S and 0 otherwise, then the estimate of $w(S)$ is given by the fraction of pages in the sample sequence that are in the index. That is,

$$\bar{w}(S) = \frac{1}{n} \sum_{i=1}^{n} I[p_i \in S].$$

Consequently, one needs to measure the fraction of pages in the sample sequence that are in the index S in order to approximate the quality of S. To achieve this, the authors pointed out that there is a requirement of two components. First, a mechanism is required for selecting pages according to w. Second, a method for testing whether a page is indexed by a search engine.

The authors adopted the approach used by Bharat and Broder [1998] in order to test whether a URL is indexed by a search engine. However, selecting pages according to a weight distribution w is significantly harder. It is also difficult to select Web pages according to the PageRank distribution. To solve this problem, Henzinger et al. [1999] proposed a sampling algorithm that provides a quality measure close to that which would be obtained using the PageRank distribution. This approach is based on the assumption that one has the means to choose a page uniformly at random. In that case, one could perform a random walk with an equilibrium distribution corresponding to the PageRank measure. At each step, the walk would either jump to a random page with probability d or follow a random link with probability $1 - d$. By executing this walk for a long period of time, one could generate a sample sequence and the pages in the sample sequence would have a distribution close to the PageRank distribution.

As pointed out by the authors, there are two problems with the implementation of such a random walk. First, no method is known for choosing a Web page uniformly at random. Second, it is not clear how many steps one would have to perform in order to remove the bias of the initial state and thereby approximate the equilibrium distribution. To solve the first problem, the authors adopted the following approach: the walk occasionally chooses a host uniformly at random from the set of hosts encountered on the walk thus far, and jumps to a page chosen uniformly at random from the set of pages discovered on that host thus far. Obviously, the equilibrium distribution of this approach does not match the PageRank distribution, since pages that have not been visited yet cannot be chosen, and pages on hosts with a small number of pages are more likely to be chosen than pages on hosts with a large number of pages. However, the authors experimentally showed [Henzinger et al. 1999] that this bias does not prevent the walk from approximating a good quality metric that show similar behavior as PageRank.

For the second problem, it is obvious that one has to start the random walk from some initial page. Hence, this introduces a bias towards pages close to the

[15] In Henzinger et al. [1999], the $(1 - d)$ term in the definition of PageRank is normalized by the total number of pages on the web T.

initial page. The authors experimentally justified that as a substantial portion of the Web is highly connected, with reasonably short paths between pair of pages, randomly walking over a small subgraph of the Web suffices to handle the initial bias [Henzinger et al. 1999].

Henzinger et al. [1999] provided some experimental results based on their random walks. First, the results demonstrate that the random walk approach does capture the intuitive notion of quality and the weight distribution appears heavily skewed towards pages user would consider useful. Second, the results compared the measured quality of several search engine indexes. For instance, comparing the quality scores of the search engine indexes according to their size, Alta Vista scored the highest under the quality metric; Excite does extremely well for a search engine of intermediate size. Similarly, comparing the average page quality of the search engine indexes, the authors found out that larger search engines sometimes have higher average page quality.

7. INFORMATION THEORETIC

The final category of Web metrics comprises metrics that measure properties related to information needs, production and consumption. Here, we consider two properties discussed by Pitkow and Pirolli [1997], namely, *desirability* and *survivability* of Web documents and *rate of change* of Web pages as discussed by Brewington and Cybenko [2000].

The desirability of a page is the probability that its information will be needed in a given time interval. Given that the frequency of page access on the WWW is approximately a negative binomial distribution, the information desirability seems to follow the Burrell Gamma-Poisson (BGP) model, which assumes that accesses to information are Poisson events and their desirability is modeled by a Poisson parameter. The logarithm of need probability satisfies a negative linear relationship with the logarithm of time since last access.

Survival analysis models the probability that a particular item will be deleted at a particular time. The survival function defined over time t is the probability that a page survives at least up to time t. That is, for a survival time T, distribution function $F(t)$, and the corresponding density function $f(t)$, the survival function is:

$$\begin{aligned} S(t) &= 1 - F(t) \\ &= P\{T > t\}. \end{aligned}$$

The *hazard rate* is defined as the probability that a page will be deleted in the next unit of time, given that it has survived to time t:

$$\Delta(t) = \frac{f(t)}{S(t)}.$$

Brewington and Cybenko [2000] developed an exponential probabilistic model for the times between individual Web page changes based on observational data on the rates of change for a large sample of Web pages. They observed pages at a rate of about 100,000 pages per day for period of over seven months, recording how and when these pages have changed. About one page in five is younger than eleven days, and half of the Web's content is younger than three months. In this context, the *age* of a Web page is defined as the difference between a downloaded page's last-modified timestamp and the time at downloading and *lifetime* is the time between changes. About one page in four is older than one year and sometimes much older than that. Inspired by this findings, the author model the changes in a single Web page as a renewal process. If $g(t)$ is the age probability density and λ is the change rate, then it can be shown that

$$g(t) = \lambda \exp(\lambda t).$$

Note that the lifetime probability density is related to the age probability density as the act of observing "the age is t units" is the same as knowing "the lifetime is no smaller than t units." Consequently, we can estimate a page's lifetime PDF, assuming an exponential distribution, using only page age observations, which can be easily obtained from the data.

The authors also modeled the age distribution for the entire Web using the joint distribution of the exponential growth rate of Web pages and the change rate λ. Using a distribution over the inverse rate $\lambda = 1/x$ and exponential growth rate parameter ϵ, it can be shown that the age probability density for the entire web is as follows:

$$g(t) = \int_0^\infty \left(\epsilon + \frac{1}{x}\right) \times \exp\left(\left(\epsilon + \frac{1}{x}\right) t\right) w(x)\, dx.$$

Note that the authors observed that the shape of the distribution $w(x)$ in the above equation roughly follows Weibull distribution Montgomery and Runger [1994], which is given by

$$w(t) = \frac{\sigma}{\delta}\left(\frac{t}{\delta}\right)^{\sigma-1} \exp\left(-(t/\delta)^\sigma\right).$$

where δ is the scale parameter and δ is a shape parameter. The shape parameter can be varied to change the shape peaked distribution to an exponential to a unimodal distribution. The scale parameter adjust the mean of the distribution. Numerically, the authors found out that the optimal values are $\epsilon = 0.00176$, $\sigma = 0.78$ and $\delta = 651.1$.

8. CONCLUSION

In this article, we have reviewed and classified some well known Web metrics. Our approach has been to consider these metrics in the context of improving Web content while intuitively explaining their origins and formulations. This analysis is fundamental to modeling the phenomena that give rise to the measurements. To our knowledge, this is the first survey that incorporates an extensive treatment of wide range of metrics and measurement functions. Nevertheless, we do not claim this survey is complete and acknowledge any omissions. We hope that this initiative would serve as a reference point for further evolution of new metrics for characterizing and quantifying information on the Web and developing the explanatory models associated with them.

REFERENCES

ALBERT, R. AND BARABASI, A. 2000. Topology of evolving networks: Local events and uncertainty. *Phys. Rev. Lett. 84*, 56–60.

ALBERT, R., JEONG, H., AND BARABASI, A. 1999. The diameter of the world wide web. *Nature 401*, 130–131.

BARABASI, A. AND ALBERT, R. 1999. Emergence of scaling in random networks. *Science 286* (Oct.), 509–512.

BARABASI, A., ALBERT, R., AND JEONG, A. 1999. Mean-field theory for scale free random networks. *Phys. A 272*, 173–187.

BARABASI, A., ALBERT, R., AND JEONG, J. 2000. Scale-free characteristics of random networks: The topology of the world wide web. *Phys. A*, *281*, 69–77.

BHARAT, K. AND BRODER, A. 1998. A technique for measuring the relative size and overlap of public web search engines. In *Proceedings of the 7th International World Wide Web Conference* (Australia, Apr.).

BREWINGTON, B. AND CYBENKO, G. 2000. How dynamic is the web? In *Proceedings of the 9th International World Wide Web Conference* (The Netherlands).

BRODER, A., GLASSMAN, S., MANASSE, M., AND ZWEIG, G. 1997. Syntactic clustering of the web. In *Proceedings of the 6th World Wide Web Conference*.

BRODER, A., KUMAR, R., MAGHOUL, F., RAGHAVAN, P., RAJAGOPALAN, S., STATA, R., TOMKINS, A., AND WIENER, J. 2000. Graph structure of the web. In *Proceedings of the 9th World Wide Web Conference*.

BOYCE, B. R., MEADOW, C. T., AND KRAFT, D. H. 1994. *Measurement in Information Science*. Academic Press Inc. Orlando, Fla.

BRIN, S. AND PAGE, L. 1998. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the 7th World Wide Web Conference*.

BORODIN, A., ROBERTS, G., ROSENTHAL, J. S., AND TSAPARAS, P. 2001. Finding authorities and hubs from link structures on the world wide web. In *Proceedings of the 10th International World Wide Web Conference* (Hong Kong).

BOTAFOGO, R., RIVLIN, E., AND SHNEIDERMAN, B. 1992. Structural analysis of hypertexts: Identifying hierarchies and useful metrics. *ACM Trans. Inf. Syst. 10*, 2 (Apr.), 142–180.

BRAY, T. 1996. Measuring the web. In *Proceedings of the 5th International World Wide Web Conference* (Paris, France. May).

CATLEDGE, L. AND PITKOW, J. 1995. Characterizing browsing strategies in the world wide web. *Comput. Netw. ISDN Syst. 27*, 6 .

CHAKRABARTI, S., DOM, B., GIBSON, D., KUMAR, R., RAGHAVAN, P., RAJAGOPALAN, S., AND TOMKINS, A.

1998a. Experiments in topic distillation. In *Proceedings of the SIGIR Workshop on Hypertext IR*.

CHAKRABARTI, S., DOM, B., RAGHAVAN, P., RAJAGOPALAN, S., GIBSON, D., AND KLEINBERG, J. 1998b. Automatic resource compilation by analyzing hyperlink structure and associated text. In *Proceedings of the 7th World Wide Web Conference*.

CHO, J., GARCIA-MOLINA, H., AND PAGE, L. 1998. Efficient crawling through url ordering. In *Proceedings of the 7th World Wide Web Conference*.

COHN, D. AND CHANG, H. 2000. Learning to probabilistically identify authorative documents. In *Proceedings of the 17th International Conference on Machine Learning* (Calif).

DEAN, J. AND HENZINGER, M. 1999. Finding related pages in the world wide web. In *Proceedings of the 8th World Wide Web Conference*.

DHYANI, D. 2001. Measuring the web: Metrics, models and methods. Master's Dissertation, School of Computer Engineering, Nanyang Technological University, Singapore.

EGGHE, L. AND ROUSSEAU, R. 1990. *Introduction to Informetrics*. Elsevier Science Publishers. Amsterdam, The Netherlands.

GIBSON, D., KLEINBERG, J., AND RAGHAVAN, P. 1998. Inferring web communities from link topology. In *Proceedings of the 9th ACM Conference on Hypertext and Hypermedia*.

HAWKING, D., CRASWELL, N., THISLEWAITE, P., AND HARMAN, D. 1999. Results and challenges in web search evaluation. In *Proceedings of the 8th World Wide Web Conference*.

HENZINGER, M., HEYDON, A., MITZENMACHER, M., AND NAJORK, M. 1999. Measuring index quality using random walks on the web. In *Proceedings of the 8th World Wide Web Conference*.

KLEINBERG, J. 1998. Authoritative sources in a hyperlinked environment. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*.

KLEINBERG, J., KUMAR, R., RAGHAVAN, P., RAJAGOPALAN, S., AND TOMKINS, A. 1999. The web as a graph: Measurements, models, and methods. In *Proceedings of the 5th International Conference on Computing and Combinatorics (COCOON)*.

KUMAR, R., RAGHAVAN, P., RAJAGOPALAN, S., AND TOMKINS, A. 1999. Trawling the web for emerging cyber-communities. In *Proceedings of the 8th World Wide Web Conference*.

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 of Information Science*.

LAWRENCE, S. AND GILES, C. L. 1998. Searching the world wide web. *Science 280* (Apr.).

LAWRENCE, S. AND GILES, C. L. 1999. Searching the web: General and scientific information access. *IEEE Commun. 37*, 1, 116–122.

LEE, D., CHUANG, H., AND SEAMONS, K. 1997. Effectiveness of document ranking and relevance feedback techniques. *IEEE Softw. 14*, 2 (Mar./Apr.), 67–75.

LEMPEL, R. AND MORAN, S. 2000. The stochastic approach for link structure analysis (SALSA) and the TKC effect. In *Proceedings of the 9th World Wide Web Conference*.

LEMPEL, R. AND SOFFER, A. 2001. PicASHOW: Pictorial authority search by hyperlinks on the web. In *Proceedings of the 10th International World Wide Web Conference* (Hong Kong).

MARCHIORI, M. 1997. The quest for correct information on the web: Hyper search engines. In *Proceedings of the 6th World Wide Web Conference*.

MONTGOMERY, D. C. AND RUNGER, G. C. 1994. *Applied Statistics and Probability for Engineers*. Wiley, New York.

MURRAY, B. H. AND MOORE, A. 2000. Sizing the internet. *White paper*. Available from http://www.cyveillance.com/web/us/downloads/Sizing_the_Internet.pdf (July).

PERKOWITZ, M. AND ETZIONI, O. 1997. Adaptive web sites: An AI challenge. In *Proceedings of the 15th International Joint Conference on Artificial Intelligence*.

PERKOWITZ, M. AND ETZIONI, O. 1998. Adaptive web sites: Automatically synthesizing web pages. In *Proceedings of the 15th National Conference on Artificial Intelligence*.

PERKOWITZ, M. AND ETZIONI, O. 1999. Towards adaptive web sites: Conceptual framework and case study. In *Proceedings of the 8th World Wide Web Conference*.

PIROLLI, P., PITKOW, J., AND RAO, R. 1996. Silk from a sow's ear: Extracting usable structures from the web. In *Proceedings of the ACM-SIGCHI Conference on Human Factors in Computing*.

PITKOW, J. 1997. In search of reliable usage data on the WWW. In *Proceedings of the 6th World Wide Web Conference*.

PITKOW, J. AND PIROLLI, P. 1997. Life, death and lawfulness on the electronic frontier. In *Proceedings of the ACM-SIGCHI Conference on Human Factors in Computing*.

RAFIEI, D. AND MENDELZON, A. 2000. What is this page known for? computing web page reputations. In *Proceedings of the 9th World Wide Web Conference*.

RECKER, M. AND PITKOW, J. 1996. Predicting document access in large multimedia repositories. *ACM Trans. Comput.-Hum. Inter. 3*, 4.

ROSS, S. 1983. *Stochastic Processes*. Wiley, New York.

SELBERG, E. AND ETZIONI, O. 1995. Multi-service search and comparison using the MetaCrawler. In *Proceedings of the 4th International World Wide Web Conference*.

SARUKKAI, R. 2000. Link prediction and path analysis using Markov chains. In *Proceedings of the 4th World Wide Web Conference*.

SNELL, L. 1998. *Introduction to Probability.* McGraw-Hill International Edition, Englewood Cliffs, N.J.

WEISS, R., VELEZ, B., SHELDON, M., NAMPREMPRE, C., SZILAGYI, P., DUDA, A., AND GIFFORD, D. 1996. Hypursuit: A hierarchical network search engine that exploits content-link hypertext clustering. In *Proceedings of the 7th ACM Conference on Hypertext*.

YAN, T., JACOBSEN, M., GARCIA-MOLINA, H., AND DAYAL, U. 1996. From user access patterns to dynamic hypertext linking. In *Proceedings of the 5th International World Wide Web Conference* (France).

YUWONO, B. AND LEE, D. 1996. Search and ranking algorithms for locating resources on the world wide web. In *Proceedings of the 12th International Conference on Data Engineering* (Mar.).

YUWONO, B., LAM, S., YING, J., AND LEE, D. 1995. A world wide web resource discovery system. In *Proceedings of the 4th International World Wide Web Conference*.

Received July 2001; revised May 2002; accepted August 2002