DATAK 658
8 February 2003 Disk used
ARTICLE IN PRESS
No. of Pages 25, DTD = 4.3.1
SPS-N, Chennai

ELSEVIER

Available at
www.ComputerScienceWeb.com
POWERED BY SCIENCE @ DIRECT®

Data & Knowledge Engineering xxx (2003) xxx–xxx

DATA & KNOWLEDGE ENGINEERING

www.elsevier.com/locate/datak

Deriving and verifying statistical distribution of a hyperlink-based Web page quality metric

Devanshu Dhyani, Sourav S. Bhowmick *, Wee-Keong Ng

School of Computer Engineering, Nanyang Technological University, Singapore 639798, Singapore

Received 19 June 2002; received in revised form 20 November 2002; accepted 22 January 2003

Abstract

The significance of modeling and measuring various attributes of the Web in part or as a whole is undeniable. Modeling information phenomena on the Web constitutes fundamental research towards an understanding that will contribute to the goal of increasing its utility. Although Web related metrics have become increasingly sophisticated, few employ models to explain their measurements. In this paper, we discuss issues related to metrics for *Web page significance*. These metrics are used for ranking the quality and relevance of Web pages in response to user needs. We focus on the problem of ascertaining the statistical distribution of some well-known hyperlink-based Web page quality metrics. Based on empirical distributions of Web page degrees, we derived analytically the probability distribution for the PageRank metric. We found out that it follows the familiar inverse polynomial law reported for Web page degrees. We verified the theoretical exercise with experimental results that suggest a highly concentrated distribution of the metric.

Keywords: Web measurement; Quality metrics; PageRank; Statistical distribution

1. Introduction

At the genesis of the WWW, few people imagined its explosive growth as they were entertained by Garrett Hardin's [17] revelation of the last days of a world that "perished peacefully, inexorably 'suffocated' as one of their prophets put it 'by their own intellectual excreta'." Figuratively, this suffocation is not very far judging from the disproportionate growth of one of mankind's

* Corresponding author.
E-mail addresses: assourav@ntu.edu.sg (S.S. Bhowmick), awkng@ntu.edu.sg (W.-K. Ng).

0169-023X/03/$ - see front matter © 2003 Published by Elsevier Science B.V.
doi:10.1016/S0169-023X(03)00034-X

greatest inventions with respect to our ability to exploit and manage it. In one of the regular surveys conducted as early as 1996 by Pitkow and Kehoe [28], a third of the users questioned reported finding it difficult to locate and organize online information. Considering the manifold expansion of the Web since then and the relatively slower improvements in search and retrieval technology, this problem can only be more severe now.

1.1. Measuring the Web

While efforts to redress similar problems for improving the capacity of the Web for serving the information needs of its users are well underway, it is possible to view the chaotic environment of the WWW from a different perspective. Given its influence and growth, the Web is itself a fascinating object of study for mathematical, sociological and commercial reasons. Our focus in this study is the *measurement* and *modeling* of interesting attributes and phenomena on the Web. This perspective does not offer solutions to the immediate problems facing Web architects, designers and developers but offers insight into fundamental aspects much like basic research in science and mathematics benefits technology.

The importance of measuring attributes of known objects in precise quantitative terms has for long been recognized as crucial for enhancing our understanding of our environment. One of the earliest attempts to make global measurements on the Web was undertaken by Bray [5]. The study answers early questions regarding the size of the Web, its connectivity, visibility of sites and the distribution of data formats. Since then, several directly observable metrics such as hit counts, click-through rates, access distributions and so on have become popular for quantifying aspects such as 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 [29] 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, a new generation of metrics that provide deeper insight on the Web as a whole and also on individual sites, is emerging.

What exactly is measurement and what are the objects of measurement on the Web? To clarify the exact meaning of some frequently used terms, we adopt the following definition [2]:

> 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 which 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.

We have identified a variety of measurable attributes on the WWW. We may classify Web related metrics with regard to their magnitudes, measurement functions and measures into the following categories based on the measurable attributes:

- *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.

- *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 [11].
- *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.
- *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.
- *Information theoretic*: Information theoretic metrics [12] 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.

1.2. Web page quality

We have discussed metrics for *usage characterization* and *information theoretic* in [11] and [12] respectively. In this paper, we focus on *certain* aspects of *significance metrics*. [1] We believe that out of the above categories of metrics, 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 [22,30] 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 [3] use a combination of relevance and quality metrics in ranking the responses to user queries. Also, page quality measures do not rely on page contents which make them convenient to ascertain and at the same time sinister "spamdexing" schemes [2] becomes relatively more difficult to implement.

Specifically, recent work in Web search such as PageRank [3], Authorities/Hubs [20] and Hyperinformation content [26] 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 judgement by p of q's content. Of course people can be (and often are) malicious: the same person can create several pages whose only purpose is to link to some other page just to make it look relevant. In this paper, we will assume a more "honest" model for the Web.

[1] A shorter version of this paper has appeared in [13].

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

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

As these measures of quality depend upon Web page in and out-degrees, knowledge of degree distribution can lead to their probability density functions. In this paper, we study the measurement of hyperlink information at a microscopic level in assessing the quality or relevance of page. We demonstrate an approach for deriving the distribution of PageRank from the empirical distributions of topological primitives. We also experimentally verify the probability distribution of PageRank. There are several reasons why this exercise is instructive. Firstly, it illustrates a generic methodology that can be extended to other hyperlink metrics. We know for instance that authority and hub weights [20] are formulated in a similar fashion to PageRank. Modeling these weights as random variables as we do here for PageRank can be the basis for characterizing these metrics statistically on a large scale. Secondly, a distribution derived theoretically from observations of more primitive determinants is likely to be more reliable than an empirically obtained one that is inextricably linked to the experimental setup. This conforms with the conventional wisdom of making measurements as fundamental as possible before deriving more comprehensive metrics. Additionally, in the case of Web hyperlink metrics such as PageRank, we avoid running computationally expensive algorithms. Finally, a theoretical distribution serves as a model that can help us predict precisely and consistently the effect of changes in certain parameters without incurring the cost of carrying out complex measurements again. One application of this has been mentioned at the outset for working out the size–quality constraints of search engines.

The rest of the paper is organized as follows: In Section 2, we discuss the problem of ascertaining the statistical distribution of a hyperlink-based Web page quality metric, i.e., PageRank. Next, in Section 3 we verify the theoretical exercise with experimental results. In Section 4, we provide an overview of some other quality metrics. Finally, we conclude by summarizing this paper.

2. Distribution of quality metrics (PageRank)

As quality measures depend upon Web page in and out-degrees, knowledge of degree distribution can lead to their probability density functions. Knowing, say the cumulative distribution of PageRank [3] for the Web F_R, one can determine the number of high-quality pages according to some threshold say r, given the size of the Web N. That is, the number of pages with PageRank greater than r can be estimated from Fig. 1 as

$$N' = N \cdot \Pr(R > r) = N(1 - F_R(r))$$

The value of N' can be useful for crawlers looking for high-quality Web pages in deciding optimum size versus quality configurations for search engine indexes. The distribution can also help

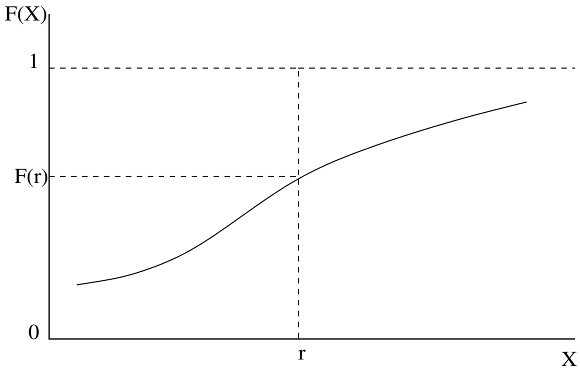


Fig. 1. Cumulative distribution.

Web crawlers give greater priority to visiting more important, high-quality pages first as done in [8].

In this section, we demonstrate an approach for deriving the distribution of PageRank from the empirical distributions of topological primitives. We first give an overview of PageRank and explain its key determinants.

2.1. PageRank

The PageRank R_i of a page i having in-degree n can be defined in terms of the PageRank of each of the n neighboring pages and their out-degrees. Let us denote by j $(1 \leqslant j \leqslant n)$, the index of neighboring pages that point to i and by X_j the out-degree of page j. Then for a fixed parameter d in $[0, 1]$ the PageRank R_i of i is given as

$$R_i = (1 - d) + d \sum_{j=1}^{n} \frac{R_j}{X_j} \qquad (1)$$

We refer to d $(0 \leqslant d \leqslant 1)$ as the damping factor for the calculation of PageRank. 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. Therefore, PageRank is a characteristic of the Web page itself—it is higher if more Web pages link to this page, as well as if these Web pages have high PageRank. Consequently, important Web pages help to make other Web pages important.

The PageRank may also be considered as the probability that a *random surfer* 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.

The PageRank is used as one component of the Google search engine [3], 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. [8] 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 [18]. However, PageRank has the problem of *Link Sink*. Link Sink occurs when page a and page b point to each other but have no links link to other pages. If page a is pointed to by an external page, during the iteration of PageRank, the loop accumulates the weight and never distributes the weight to other pages. This causes the oscillation of the algorithm and the algorithm cannot converge.

We derive the distribution of a simplified version of PageRank, ignoring the recurrent relationship of R_i with the PageRank of other pages R_j and assuming the formulation to be

$$R_i = (1 - d) + d\sum_{j=1}^{N_i} \frac{1}{X_j} \tag{2}$$

Computationally, the determination of PageRank for a graph of k pages can be seen as equivalent to the steady state solution ($n \to \infty$) of following matrix product relationship:

$$\begin{pmatrix} R_1^{n+1} \\ R_2^{n+1} \\ \vdots \\ R_k^{n+1} \end{pmatrix} = \begin{pmatrix} 1-d \\ 1-d \\ \vdots \\ 1-d \end{pmatrix} + d \begin{pmatrix} \frac{1}{x_{11}} & \cdots & \frac{1}{x_{i1}} & \cdots & \frac{1}{x_{k1}} \\ \frac{1}{x_{12}} & \cdots & \frac{1}{x_{i2}} & \cdots & \frac{1}{x_{k2}} \\ \vdots & & \vdots & & \vdots \\ \frac{1}{x_{1k}} & \cdots & \frac{1}{x_{ik}} & \cdots & \frac{1}{x_{kk}} \end{pmatrix} \cdot \begin{pmatrix} R_1^{n} \\ R_2^{n} \\ \vdots \\ R_k^{n} \end{pmatrix}$$

where R_i^n denotes the PageRank of page i at the nth iteration and x_{ij} the number of links from page i to j. If there are no outgoing links from page i to j, i.e., $x_{ij} = 0$, then the corresponding entry in the matrix $(1/x_{ij})$ is set to zero. Repeated multiplication of inverse out-degree matrix with the PageRank vector yields the dominant eigenvector of the latter. PageRank can thus be seen as the stationary probability distribution over pages induced by a random walk on the Web, that is, it represents the proportion of time a "random surfer" can be expected to spend visiting a page. It is clear that the steady state distribution of the PageRank vector (R_i^n) depends entirely on the value of d and the right hand vector of inverse out-degrees. Our simplification of Eq. (2) aims at finding the distribution of this vector at $n = 0$. The distribution of (R_i^1) gives us an idea of the steady state distribution at $n \to \infty$ which can itself be obtained by applying the above computation on the initial distribution. We further assume that initially PageRank is uniformly distributed, that is, $R_i^0 = 1/k$ for all i $(1 \leqslant i \leqslant k)$.

We interpret R_i, X_j and N_i in Eq. (2) as random variables denoting PageRank of i, the out-degree of j and the in degree of i respectively. Although both X_j and N_i are known to have the same distribution, X_j is continuous while N_i is discrete. It is clear that R_i for all values of i are identically distributed. The same holds for the in- and out-degrees denoted by X_j and N_i. We therefore represent the common probability densities of these sets of random variables as $f_R(r)$, $f_X(x)$ and $f_N(n)$ respectively. The problem now is to find the density $f_R(r)$ given the relationship of R_i with X_j and N_i as represented by Eq. (2).

2.2. The Lotka density

The derivation of the distribution of PageRank is based on observations of distribution of Web page degrees. These measurements carried out on Web graphs by Broder et al. [1] and Kleinberg et al. [21] have been reported to follow the well-known *Lotka distribution*. The Lotka density is given as

$$f_X(x) = \begin{cases} \dfrac{C}{x^{\alpha}} & \text{if } x \geqslant 1 \\ 0 & \text{otherwise} \end{cases}$$

where $\alpha \approx 2$ and C is a constant. The Lotka distribution is a frequently studied phenomenon in the field of bibliometrics for citations in academic literature [14]. In our derivation, we invoke both the continuous and discrete versions of this law. Here we distinguish between the two and examine the implications of each. If we interpret X as a continuous random variable, the constant C is found using the fact that the area under a probability density curve sums to unity. That is,

$$\int_{-\infty}^{\infty} f_X(x)\,\mathrm{d}x = 1$$

Applying this to the continuous Lotka density above, we have

$$\int_{1}^{\infty} \frac{C}{x^2}\,\mathrm{d}x = 1 \qquad (3)$$

Solving this we obtain

$$C = 1$$

The continuous version of Lotka's law can then simply be stated as follows:

$$f_X(x) = \begin{cases} \dfrac{1}{x^{2}} & \text{if } x \geqslant 1 \\ 0 & \text{otherwise} \end{cases}$$

In the case of the discrete version, the integral of Eq. (3) changes to a discrete summation, hence,

$$\sum_{x=1}^{\infty} \frac{C}{x^2} = 1$$

If we factor the constant C from the summation on the left, we are left with the sum $\sum_{x=1}^{\infty} x^{-2}$ which is the well-known *Riemann zeta function* $\zeta(2)$. Several analytical methods exist for computing the zeta function. Here we use the following general definition for even arguments, i.e., $n \equiv 2k$

$$\zeta(n) = \frac{2^{n-1}|B_n|\pi^n}{n!}$$

where $|B_n|$ is a Bernoulli number. Given $B_2 = 1/6$, we have for $n = 2$

$$\zeta(2) = \frac{\pi^2}{6}$$

Substituting this we obtain the expression for C as

$$C = \frac{1}{\zeta(2)} = \frac{6}{\pi^2} \approx 0.608$$

Although the in- and out-degree distributions on the WWW have been discovered principally as discrete distributions, we apply the continuous approach in examining the relationship between degree and PageRank distributions because the apparatus of infinitesimal calculus makes the mathematical formulation easier. Indeed, as the number of pages being considered increases, the differences between the continuous and discrete approaches become insignificant.

To approximate the Lotka distribution function $F_X(x) = \Pr(X \leqslant x)$, we integrate the density function $f_X(\cdot)$ within the continuous range $(1, \infty)$

$$F_X(x) = \int_1^x f_X(z)\,\mathrm{d}z = \int_1^x \frac{1}{z^2}\,\mathrm{d}z = 1 - \frac{1}{x}$$

2.3. The PageRank distribution

The non-recurrent definition of PageRank in Eq. (2) may be viewed as a composition of three primitive functions of random variables enumerated below.

(1) *The inverse of individual out-degrees of pages*: The individual out-degrees are independent identically distributed random variables, X_j with the index j ranging from 1 to the in-degree of the page being considered. If we represent the inverse of the out-degree of the jth neighboring page as a random variable Y_j then,

$$Y_j = \frac{1}{X_j} \tag{4}$$

It is known that the density of out-degrees, $f_X(x)$ is the Lotka function introduced earlier. We denote the density function of Y_j as $f_Y(y)$, since Y_j is identically distributed for all j.

(2) *The sum of out-degree inverses*: We denote by a random variable Z_i, the sum of out-degree inverses Y_j. That is,

$$Z_i = \sum_{j=1}^{N_i} Y_j \tag{5}$$

The upper limit to the sum is itself a random variable denoting the out-degree of the page in question. Fortunately, this random variable N_i, the out-degree of page i, is Lotka distributed. We must note however, that N_i is necessarily a discrete random variable as must any index to a discrete summation. Thus, N_i has the probability density obtained earlier for the Lotka distribution

$$f_N(n) = \begin{cases} \frac{6}{\pi^2}\frac{1}{n^2} & \text{if } n \geqslant 1 \\ 0 & \text{otherwise} \end{cases}$$

(3) Finally, the PageRank function of Eq. (2) can be expressed as a linear function of the random sum Z_i above as

$$R_i = (1 - d) + dZ_i \tag{6}$$

We now determine the densities of the random variables Y_i, Z_i and R_i introduced above. Consider Y_i, the inverse of X_i that represents the out-degree of page i. We first note that Y_i is a strictly decreasing [3] function in the range of X_i (with the latter defined on positive values only). The probability distribution of Y_j can be expressed in terms of the distribution of X_j as follows:

$$\begin{aligned} F_Y(y) &= \Pr(Y_j \leqslant y) \\ &= \Pr\left(\frac{1}{X_j} \leqslant y\right) \\ &= \Pr\left(X_j \geqslant \frac{1}{y}\right) \text{ since } \frac{1}{X_j} \text{ is strictly decreasing} \\ &= 1 - \Pr\left(X_j < \frac{1}{y}\right) \\ &= 1 - F_X\left(\frac{1}{y}\right) \end{aligned}$$

Differentiating the above form to convert to probability densities, we have

$$\begin{aligned} f_Y(y) &= \frac{\mathrm{d}}{\mathrm{d}y} F_Y(y) \\ &= -f_X\left(\frac{1}{y}\right) \frac{\mathrm{d}}{\mathrm{d}y}\left(\frac{1}{y}\right) \\ &= \frac{1}{y^2} f_X\left(\frac{1}{y}\right) \end{aligned}$$

Substituting the Lotka continuous density function for $f_X(\cdot)$ to the above result and applying the range $[1, \infty]$ to the argument $1/y$, we obtain the uniform density for Y_j over the converted range $(0, 1]$

$$f_Y(y) = \begin{cases} 1 & \text{if } 0 < y \leqslant 1 \\ 0 & \text{otherwise} \end{cases}$$

The above result, that the inverse of a Lotka distributed random variable has a uniform distribution is an interesting coincident. Intuitively it implies that even though the probability that the out-degree of a page is in a given range follows an inverse square law, the probability of out-degree inverse is uniformly distributed, i.e., independent of the value of Y_j.

The sum of out-degree inverses Z_i given by $\sum_{j=1}^{n} Y_j$ has a variable number of terms equal to n, the number of pages that point to i or the out-degree of i. We model the limit of the summation itself as a discrete random variable N_i. Such a sum is commonly referred to as a *random sum*. As noted earlier, N_i is Lotka distributed and similar to X_i except that the distribution here is the discrete version of the Lotka function. Note that Z_i for all values of i are identically distributed so

[3] A function ϕ is strictly decreasing if, $\phi(x_1) > \phi(x_2)$ when $x_1 < x_2$ for any two values x_1 and x_2 in its domain.

the common density can be denoted $f_Z(z)$ as done earlier for the random variables X_i and Y_i. We first find the density of Z_i conditioned on the summation limit N_i, that is $f_{Z|N}(z|n)$.

A sum of random variables has a density which is the convolution of the densities of individual random variables. For n identically distributed summands, this specializes to the n-fold convolution of their common density, in this case $f_Y(y)$. This n-fold convolution $f_Y^{(n)}(y)$ is defined recursively as follows:

$$f_Y^{(1)}(y) = f_Y(y) \text{ and}$$
$$f_Y^{(n)}(y) = \int_{-\infty}^{\infty} f_Y^{(n-1)}(y-u) f_Y(u)\,\mathrm{d}u \quad \text{for } n > 1$$

The above definition can be applied to $f_Y(y)$ derived earlier to obtain the following formula for the density of the random sum Z_i conditioned upon the out-degree of page i. Fig. 2 shows the n-fold convolutions of the uniform density for several values of n. Observe that for higher values of n, the curve flattens out resembling a normal distribution. This is predicted by the central limit theorem which states that the sum of n independent random variables tends to a uniform distribution as $n \to \infty$

$$f_{Z|N}(z|n) = f_Y^{(n)}(y)$$

where

$$f_Y^{(n)}(y) = \begin{cases} \dfrac{1}{(n-1)!} \displaystyle\sum_{j=0}^{x} (-1)^j \binom{n}{j} (x-j)^{n-1} & \text{if } 0 < x < n \\ 0 & \text{otherwise} \end{cases}$$

By the law of total probability, the continuous marginal density of the sum Z_i can be found as

$$f_Z(z) = \sum_{n=1}^{\infty} f_{Z|N}^{(n)}(z|n) \cdot f_N(n)$$

Substituting for the above expression for $f_{Z|N}(z|n)$ and the discrete Lotka density for $f_N(n)$ we have

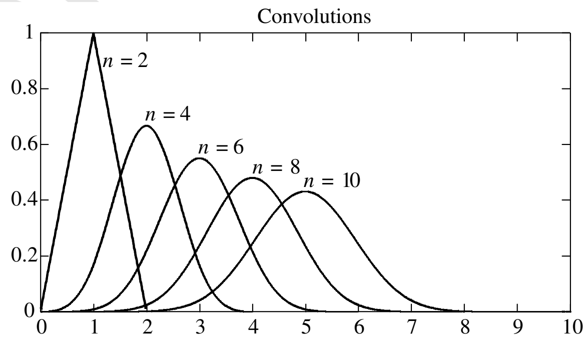


Fig. 2. Convolutions of n uniform densities.

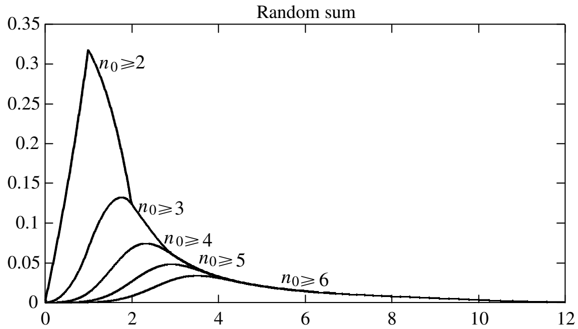


Fig. 3. Density of a sum of uniformly distributed random variables. Each curve represents the density function given by $\sum_{n=n_0}^{\infty} f_{Z|N}(z|n)f_N(n)$. For computation purposes we approximated the upper limit to be $n = 20$.

$$f_Z(z) = \begin{cases} \dfrac{6}{\pi^2} \displaystyle\sum_{n=1}^{\infty} \frac{1}{n^2(n-1)!} \sum_{j=0}^{x} (-1)^j \binom{n}{j} (x-j)^{n-1} & \text{if } 0 < x < n \\ 0 & \text{otherwise} \end{cases}$$

It is difficult to simplify the above summation since the inner sum does not have a closed form. Fig. 3 shows the approximate curve for the density $f_Z(z)$. The parameter n_0 represents the *minimum in-degree* considered for computing the random sum. Notice that the sharp peak for the curve $n_0 = 2$ occurs due to the influence of the 2-fold convolution of Fig. 2. For higher starting values of n, signifying more densely connected pages the curve becomes more even.

Finally, to derive the density $f_R(r)$ of a linear function of Z_i, we adopt a similar approach as before for finding the density of out-degree inverse, except that the function $\phi(Z_i) = 1 - d + dZ_i$ is strictly increasing. We have from Eq. (6)

$$\begin{aligned} F_R(r) &= \Pr(R_i \leqslant r) \\ &= \Pr(1 - d + dZ_i \leqslant r) \\ &= \Pr\left(Z_i \leqslant \frac{r-1}{d} + 1\right) \text{ since } \phi \text{ is strictly increasing} \\ &= F_Z\left(\frac{r-1}{d} + 1\right) \end{aligned} \tag{7}$$

and differentiating to obtain probability density,

$$\begin{aligned} f_R(r) &= \frac{\mathrm{d}}{\mathrm{d}r} F_R(r) \\ &= f_Z(\phi^{-1}(r)) \frac{\mathrm{d}}{\mathrm{d}r}(\phi^{-1}(r)) \\ &= \frac{1}{d} f_Z\left(\frac{r-1}{d} + 1\right) \end{aligned} \tag{8}$$

Fig. 4 shows the unnormalized PageRank distribution $F_R(r)$ denoting the probability $\Pr(R < r)$. For $n_0 = 2$, nearly two-thirds pages have a PageRank within 10% of its total range which means

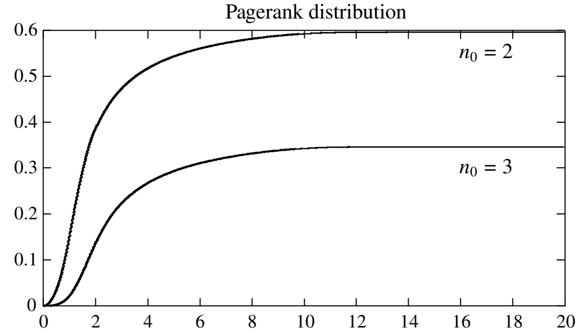


Fig. 4. Derived PageRank probability distribution $F_R(r)$ with parameter $d = 0.5$ for two values of minimum in-degree n_0.

that PageRank follows a highly concentrated distribution as do Web page in- and out-degrees. This confirms our earlier conjecture that PageRank distribution is affected by the degree distributions.

3. Experiments

We now discuss experiments conducted to verify the theoretical exercise of the previous section. We do so in two parts—first we verify that Web page out-degrees follow a Lotka distribution and then compare the derived theoretical distribution for PageRank with the observed distribution of the metric, computed over a given sample of pages.

Our experiments required large crawls [4] of the WWW and were undoubtedly limited by the scale of available computing resources, primarily memory and network bandwidth. Our crawls were different from those performed by a typical search engine in that they did not download pages for storage, but merely obtained the URL and hyperlink connectivity of visited pages. The graph structure of a set of pages is sufficient to measure degree distribution and compute PageRanks. Despite the apparently modest requirements, crawls of over a thousand pages could not be handled without memory overflows due to the additional requirements of Crawler's queues and PageRank data structures. With some modifications to the original PageRank algorithm discussed in this section, we doubled the capacity to nearly two thousand pages. Let us first briefly consider the setup of the crawler used for the experiments.

3.1. Crawler configuration

Web crawls for our experiments were performed using a Java-based toolkit for developing customized crawlers, known as SPHINX (for Site-oriented Processors for HTML Information) [25].

[4] Crawling the Web involves a *crawler* or *robot* that autonomously visits Web sites and downloads pages for indexing or other purposes.

Table 1
Web site graph structure

Source	Destination
URL1	URL2, URL3, URL4
URL2	URL5, URL6
URL3	URL7, URL8
⋮	⋮

SPHINX is especially suited for the kind of crawls we perform on the Web since it allows developers to customize the crawler to the needs of specific applications. SPHINX also supports personal crawls, where a specific task of interest to perhaps a single user, is required to be performed only a limited number of times.

The WWW is regarded by SPHINX as a directed graph whose pages and links are represented by separate objects. To create a custom crawler, the developer simply extends generic class for crawlers and overrides methods that determine which pages are to be visited and how a page should be processed when visited. The implementation of the generic class uses multiple threads to retrieve pages and places them in a queue of pages approved for visit by the user customized method. Upon visiting a page, the crawler performs a user specified set of instructions and decides whether its links are to be enqueued for visiting. In our case, we merely add the page URL to a list for creating the hyperlink graph and discard its contents.

Storage for a Web graph has two components—URLS and links. The graph is stored as a list, each of whose records contain a page URL and a list of link URLS representing the outgoing links. This structure is shown in Table 1. As we mentioned earlier, large sized graphs cannot be stored in their complete form in memory. An alternative to conserve memory is to partition the graph into blocks for storage in secondary memory and modify the PageRank computation algorithm to handle the new data organization. However, a completely fragmented graph proves too expensive at a later stage, when *graph completion* (explained later) is performed. We therefore maintain two data structures, one for the graph and another as described previously, containing a list of all visited URLS. Web pages frequently have high out-degrees; pages with hundreds of outgoing links are not uncommon. This implies that the latter data structure (containing visited URLS) is only a small proportion of the overall graph size. Consequently, holding the URL list consistently in memory is inexpensive in terms of memory usage. In our experiments, we used a block size of 50 pages. The block size parameter has an impact on the time versus memory tradeoff similar to that of page size in operating systems. The larger the block size, the smaller the number of blocks to be swapped in and out of memory but the greater the demand on main memory. With this organization of crawl data we were able to store a graph of more than 2000 pages.

3.2. Degree measurements

The first series of experiments were conducted to confirm the earlier reported distributions [1,21] of Web page in- and out-degrees. In recent work, Kleinberg et al. [21] reported that both in- and out-degrees of Web pages have *power law* distributions. A power law on positive integers describes the probability of value i as proportional to $1/i^k$ for a small positive number k. The value

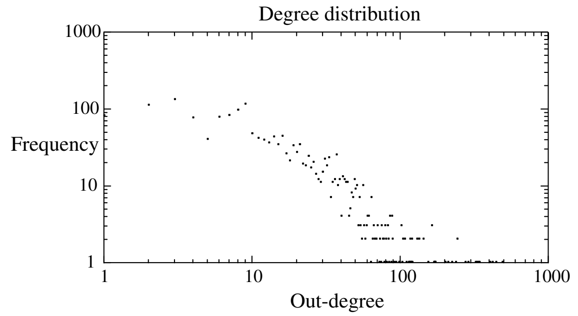


Fig. 5. Log–log plot of out-degree distribution of 2024 Web pages in the NTU domain.

of k was empirically determined to be approximately 2, giving rise to the specialization called *Lotka's law*. It was further found that the power law appears as both macroscopic phenomenon for the entire Web and as a microscopic phenomenon for individual Web sites.

To test the power law phenomenon which serves as our own starting hypothesis in the theoretical determination of PageRank distribution, we conducted experiments on a crawl of over 2000 pages. A log–log plot of out-degree distributions is shown in Fig. 5. This plot appears linear as expected for a power law distribution.

Kleinberg et al. further reported that the average out-degree is approximately 7. While this number may sound intuitively correct, we must bear in mind that due to the high variance of the Lotka distribution, the mean can be misleading. We therefore characterize its exact form by fitting the degree distribution data to the analytical Lotka function. Fig. 6 shows the results of this exercise. The exponent β of the fitted Lotka distribution $\alpha x^{-\beta}$ is close to value reported elsewhere [1,21].

3.3. PageRank measurements

We now discuss our experiments on measuring PageRank for a sample set of Web pages and comparing the empirical distribution with the one derived earlier. The computation of PageRank

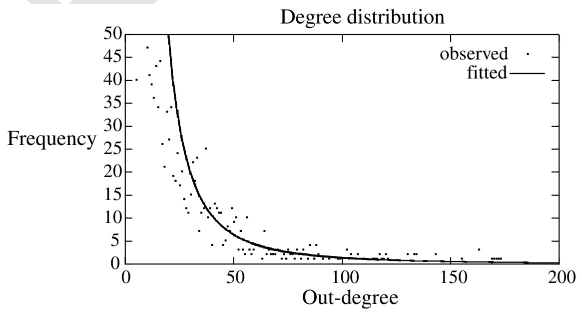


Fig. 6. Fitted out-degree distribution for a crawl size of 2024 pages. The analytical form for fitted curve is $y = \alpha x^{-\beta}$ where $\alpha = 40754.1$ and $\beta = 2.24$. The reduced χ^2 variable was 8.22. Asymptotic standard errors for α and β were 72.5% and 8.9% respectively.

is essentially an eigenvector problem as outlined in the previous section. We briefly explain the formulation before describing two algorithms for computing it efficiently and reporting the results.

The motivation behind PageRank is that a link from a page i to page j is an implicit conferral of authority to j. The amount of importance conferred is assumed to be in direct proportion to the authority of the source i and in inverse proportion to the number of outgoing links in i. Thus, if we denote by $N_{\text{in}}(i)$ and $N_{\text{out}}(i)$ the set of pages in the neighborhood of i via incoming and outgoing links respectively, we perform the following iterative computation to obtain the rank of j:

$$R_j^{n+1} = \sum_{i \in N_{\text{in}}(j)} \frac{R_i^n}{|N_{\text{out}}(i)|}$$

The initial ranks R_i^0 are set to $1/N$ where N is the total number of pages being considered. The above iterations are continued until the values R_j stabilize. The matrix equivalent of the above calculation would be

$$R^n = M \cdot R^{n-1} \tag{9}$$

where $R^n = (R_i^n)$ is the vector of PageRank and $M = (m_{ij})$ is the matrix of inverse out-degrees. This calculation leads to the principal eigenvector of R, that is, a value of R^* that satisfies the relation $R^* = M \cdot R^*$. To detect this equality within a certain error estimate we compute the *residual error* at some intermediate iteration i as $\Delta = M \cdot R^i - R^i$. But from Eq. (9) the minuend is R^{i+1}, so the residual error is simply $R^{i+1} - R^i$ or the vector distance between successive rank vectors. After adequate number of iterations we may expect the residual error to approach zero. For computation purposes we can estimate R^* as the value for which the residual error is below a specified estimation threshold.

The calculation of Eq. (9) are not guaranteed to converge. An important condition for convergence is that each node in the graph have at least one outgoing link; i.e., the graph should be *connected*. There are two options for ensuring graph connectivity. We can iteratively remove nodes with zero out-degree or add a complete set of outgoing links to any node with zero out-degree. The tradeoff between the two alternatives is that of time versus memory. The first one saves memory by reducing the graph size but is time consuming since the procedure must be performed iteratively to avoid "dangling" links (we elaborate on this later). The latter approach increases memory requirements for storing the graph. For each node without children a number of links equal to one less than the graph size must be added. Due to limited resources, we have chosen the memory conserving option of removing unconnected nodes. For convergence purposes, the two are equivalent. To avoid local minimas during convergence, we add a damping factor d in the propagation of rank in Eq. (9) as follows:

$$R^n = \left(\frac{1-d}{N}\right)I + dM \cdot R^{n-1} \tag{10}$$

where I is the unit vector of size N, the total number of pages. PageRank can in fact be *personalized* by initializing the rank vector R such that certain categories of pages experience higher rank propagation. Thus, we have in Eq. (10) the final form of PageRank as stated in the previous section.

3.3.1. Naive algorithm

Let us consider a simple implementation of PageRank regardless of memory limitations as introduced by Haveliwala [19]. Our purpose for doing so is to firstly understand thoroughly the matrix vector multiplication in the context of PageRank and secondly to discuss some of the common steps performed irrespective of the algorithm used.

As mentioned earlier, a necessary condition for convergence of the PageRank vector to a unique value is graph connectivity. We ensure graph connectivity through *node removal* by iterating through the following two-step procedure until no more pages can be removed from the graph:

(1) Remove pages with zero out-degree; that is, remove from the structure of Table 1 any page with no outgoing links. The page is removed from both the graph as well as the list of URLS.
(2) Remove *dangling links* or links that point to a page that is not present in the graph. This can be done by testing for page existence in the same structure or in the list of URLS.

To compute PageRanks, we create two floating point arrays of size N, the number of pages in the graph, representing the rank vectors at successive iterations called *Succ* and *Prev*. That is, if *Prev* is the rank vector at iteration i, then the ranks for the next iteration $i+1$ are held in *Succ*. Each entry in the PageRank vector, *Prev* is initialized to $1/N$. Let us assume the vectors *Prev* and *Succ* can be indexed by URLS from the graph structure of Table 1. For instance, *Prev*[URL1] is the PageRank of the page identified by URL1 at the current iteration. We represent the graph of Table 1 with a vector *Source* for page URLS and a commonly indexed matrix *Destination* for link URLS. Thus, $Source[i]$ denotes the URL of a page indexed i and $Destination[i][j]$ is the URL of the jth link in page i. The out-degree of a page i is simply the size of the ith row in *Destination* or $|Destination[i]|$.

Fig. 7 shows the naive algorithm for computing PageRank. At each iteration we propagate the rank contribution of each page to its neighboring pages as a function of its PageRank and out-degree and normalize the *Succ* vector by the damping factor. The iterations are continued until the rank vectors are seen to converge to a stable value. As presented earlier this condition is detected when the distance between the *Prev* and *Succ* vectors, called *residual error*, falls below a

```
Input: Web graph Source and Destination vectors.
Output: Final Pagerank vector Succ

∀s Prev[s] = 1/N
while(residual > ε) {
    ∀d Succ[d] = 0
    for i = 1...N {
        source = Source[i]
        for j = 1...|Destination[i]|
            Succ[j] = Succ[j] + Prev[source]/|Destination[i]|
    }
    ∀s Succ[s] = (1 − d)/N + d × Succ[d]  /* damping */
    residual = ||Succ − Prev||  /* compute only every few iterations */
    Prev = Succ
}
```

Fig. 7. Naive PageRank algorithm.

certain threshold ϵ in Fig. 7. Alternatively, we can also stop the calculation of PageRank after a specified number of iterations have been performed.

3.3.2. Block-based algorithm

The algorithm of Fig. 7 computes PageRank functionally but is not scalable to graph size beyond a few hundred pages because it holds the PageRank as well as graph structures (Table 1) in main memory throughout the computation. We observe that to propagate the contribution of a page we only need to know the URLS it links to. Hence, it is unnecessary to maintain the entire graph in memory while propagating ranks. This leads to a modified *block*-based algorithm of Fig. 8.

The block-based algorithm is largely similar to the naive algorithm except that the *Source* and *Destination* arrays are stored on disk and read block by block for propagating ranks. In Fig. 8, β denotes the number of blocks. The page URLS of each block are read into the *Source* vector while the links are read into the *Destination* matrix from the disk. In this scheme, the total size of the resident graph structure never exceeds $2N/\beta$ records, signifying a substantial conservation of memory. Using the block-based strategy with a block size of 50 pages, we computed the PageRank of over 2000 pages ($\beta = 40$) with the objective of confirming the theoretically derived distributions of the previous section. Following graph completion, the computation converged with residual error below a threshold of 10^{-4}. We used a damping factor of 0.5. A log–log frequency plot of PageRank values is shown in Fig. 9. The linearity of this plot, as for Web page outdegrees in Fig. 5, suggests that PageRanks also follow a power law distribution. We confirmed this intuition by fitting the observed PageRank distribution with a power law function shown in Fig. 10.

Finally, we present in Fig. 11 the comparison of the *cumulative* PageRank distribution with its theoretically derived counterpart $F_R(r)$ from the previous section (see Eq. (7)). As mentioned earlier, the theoretical expression for PageRank distribution employs infinite convolutions of uniform densities. For computation purposes, we approximated this with a limit of 20-fold convolution. The derived cumulative distribution was obtained by integrating the area under the density curve of Eq. (8) using the *trapeziod rule*. We see that the observed distribution of Page-

```
Input: Web graph Source and Destination vectors.
Output: Final Pagerank vector Succ

∀s Prev[s] = 1/N
while(residual > ε) {
    ∀d Succ[d] = 0
    for k = 1...β {
        Read block k into Source
        for i = 1...N_k {
            source = Source[i]
            for j = 1...|Destination[i]|
                Succ[j] = Succ[j] + Prev[source]/|Destination[i]|
        }
    }
    ∀s Succ[s] = (1 − d)/N + d × Succ[d]  /* damping */
    residual = ||Succ − Prev||  /* compute only every few iterations */
    Prev = Succ
}
```

Fig. 8. Block-based PageRank algorithm.

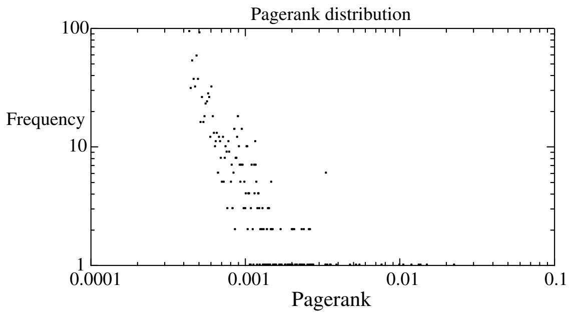


Fig. 9. Log–log plot of PageRank distribution of 2000 Web pages in the NTU domain.

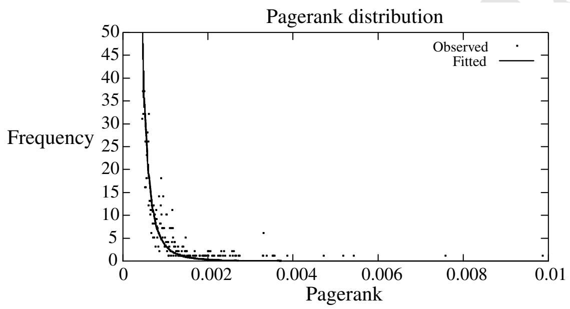


Fig. 10. PageRank distribution for 1889 pages in the NTU domain fitted to the analytical curve $y = \alpha x^{-\beta}$ where $\alpha = 1.7 \times 10^{-10}$, $\beta = 3.43$. Goodness of fit parameters: $\chi^2 = 47.78$, asymptotic standard errors for α and β were 180.9%, 6.84% respectively.

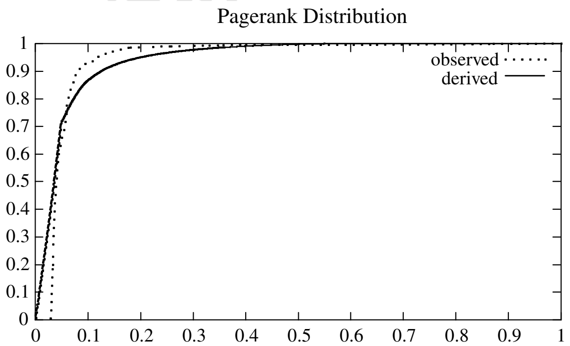


Fig. 11. Theoretical probability distribution function of PageRank (computed with 20-fold convolution) and the observed normalized cumulative distribution of NTU PageRank for a crawl of 2026 pages.

Rank compares reasonably with the distribution of Eq. (7), derived from out-degree distributions. As postulated in the theoretical derivation, out-degree distributions have a significant impact on the distribution of PageRank. The power law distribution of our-degrees is attributed by

Kleinberg et al. [21] to a *random copying process* of Web page creation. When a user encounters pages that she finds interesting, she includes links to them in her page. This process leads to the creation of locally dense subgraphs around the topic of interest. Since out-degree is the fundamental input for computing PageRank it is intuitive to suggest that PageRank will itself follow a power law distribution.

4. Some related Web page quality metrics

The preceding sections illustrate a generic methodology that can be extended to other hyperlink metrics which are formulated in a similar fashion to PageRank. In this section we give a brief overview of some of these metrics for Web page quality. These metrics rely on the link structure of the Web to rank pages. 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. Specifically, we focus our discussion on the notion of *hubs* and *authorities* in Kleinberg's algorithm [20] and some of the variants of this algorithm and the PageRank. Modeling the hub and authority weights as random variables as we do here for PageRank can be the basis for characterizing these metrics statistically on a large scale.

4.1. 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 [20]. He suggested that Web page importance should depend on the search query being performed. This model is based on a mutually reinforcing 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 which link to pages in the Root Set, and also pages which 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 $A_{ij} = 0$.

Authority and hub weights can be used to enhance Web search by identifying a small set of high-quality pages on a broad topic [6,7]. Pages related to a given page p can be found by finding the top authorities and hubs among pages in the vicinity to p [10]. The same algorithm has also been used for finding densely linked communities of hubs and authorities [15].

One of the limitations of Kleinberg's [20] *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 [9]. 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 on-line newspaper articles in the Cincinnati Enquirer which discuss the Jacksonville Jaguars football team, and all link to the same Cincinnati Enquirer service pages.

An important difference between Kleinberg's algorithm and PageRank is that PageRank is query-independent, whereas hubs and authorities depends heavily on the subject we are interested in. In PageRank all pages on the Web are ranked on their *intrinsic* value, regardless of topic. Hence, whenever a query is made, PageRank must be combined with query-specific measures to determine the relative importance in a given context. On the other hand, instead of globally ranking pages, hubs and authorities assign ranks that are specific to the query we are interested in.

4.2. 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 [27]. In the one-level influence propagation model of PageRank, a surfer performing a random walk may jump to a page chosen uniformly at random with probability *d or* follow an outgoing link from the current page. Rafiei and Mendelzon 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 [20], outgoing links can be followed *directly* from the current page p, or *indirectly* through a random page q that has a link to p.

4.3. SALSA

Lempel and Moran [23] 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 [5] 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 which 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).

SALSA does not have the same *mutually reinforcing structure* that Kleinberg's algorithm does. The relative authority of site within a connected component is determined from local links, not

[5] According to [27], a *random walk* on a set of states $S = \{s_1, s_2, \dots, s_n\}$, 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.

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 [4]. 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 [6] 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.

4.4. Approach of Borodin et al.

Borodin et al. proposed a set of algorithms for hypertext link analysis in [4]. We highlight some of these algorithms here. The authors proposed a series of algorithm which 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 [4]. 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 [4].

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.

[6] A *Markov chain* is simply a sequence of state distribution vectors at successive time intervals, i.e., $\langle \Pi^0, \Pi^1, \ldots, \Pi^n \rangle$. A Markov chain is *ergodic* if it is possible to go from every state to every other state in one or more transitions.

The *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 *i*th 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 network approach, namely, *Bayesian algorithm* and *simplified Bayesian algorithm*, as opposed to the more common algebraic/graph theoretic approach. They experimentally verified that the *simplified Bayesian algorithm* is almost identical to the SALSA algorithm and have at least 80% overlap on all queries. 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.

4.5. PicASHOW

PicASHOW [24] 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. The basic premise is that a page p displays (or links to) an image when the author of p considers the image to be of value to the viewers of the page. It does not require any image analysis whatsoever and no creation of taxonomies for pre-classification 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 which are co-contained in pages are likely to be related to the same topic. (2) Images which 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 which are contained in authoritative pages on topic t are good candidates to be quality images on that topic.

PicASHOW's analysis of the link structure enables it to retrieve authoritative images as well as to identify image containers and image hubs. The authors define these as Web pages that are rich in relevant images, or from which many images are readily accessible. Results in this work demonstrate that PicASHOW, while relying almost exclusively on link analysis, compares well with dedicated WWW image retrieval systems. The authors conclude that link analysis, a bona-fide effective technique for Web page search, can improve the performance of Web image retrieval, as well as extend its definition to include the retrieval of *image hubs* and *containers*.

5. Conclusions and future work

The emergence of the World Wide Web as an unparalleled medium of creating, sharing and disseminating information at low cost has evoked heightened research activity aimed at improving information systems aspects such as search and retrieval effectiveness. However, its rapid, unrelenting growth and influence on society and business also makes it an interesting object of fundamental research. Specifically, it necessitates the measurement and modeling of various aspects.

Apart from being of obvious theoretical interest, models reproduce essential aspects of the WWW that can improve the efficacy of Web information systems.

In this paper we treated the problem of ascertaining the statistical distribution of some well-known hyperlink-based Web page quality metrics. Based on empirical distributions of Web page degrees, we derived analytically the probability distribution for the PageRank metric and found it to follow the familiar inverse polynomial law reported for Web page degrees. We verified the theoretical exercise with experimental results that suggest a highly concentrated distribution of the metric. Our work on distributions of hyperlink metrics can be extended by conducting similar exercises for other types of significance metrics for both quality and relevance. It is not clear whether the high concentration depicted by hyperlink metrics is a special consequence of the power law behavior of Web page degrees or the manifestation of a general regularity in Web page significance, irrespective of the metric in particular.

Another interesting area which we would like to explore is the affect of the human factor on the probability distribution for the quality metrics. Our work in this paper is based on the notion of Web page quality as discussed in [3,8,20]. That is, Web page quality is measured without considering the human factor comprehensively. To measure the usefulness of a page several other measures could be combined with the PageRank (or any quality metrics in general):

- Search engine leads (whether users actually select the page when presented as a query result).
- Number of visits (referral and popularity).
- Time spent by visitors (usefulness of content).
- Interaction with dynamic content and user interface (implies utility).
- Local navigation (implies user's interest through intent to explore further).

However, each of the above is vulnerable to misinterpretation, for example a large number of visits to a Web page occur for the same reason as a high PageRank, i.e., high visibility. Similarly, time spent is influenced not just because how interesting the content is but also by Web page layout, design and readability.

6. Uncited reference

[16]

References

[1] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, J. Wiener, Graph structure of the Web, in: Proceedings of the Ninth World Wide Web Conference, Amsterdam, Netherlands, May 2000.

[2] B.R. Boyce, C.T. Meadow, D.H. Kraft, Measurement in Information Science, Academic Press Inc., 1994.

[3] S. Brin, L. Page. The anatomy of a large-scale hypertextual Web search engine, in: Proceedings of the Seventh World Wide Web Conference, Brisbane, Australia, April 1998.

[4] A. Borodin, G. Roberts, J.S. Rosenthal, P. Tsaparas, Finding authorities and hubs from link structures on the World Wide Web, in: Proceedings of the Tenth International World Wide Web Conference, Hong Kong, 2001.

[5] T. Bray, Measuring the Web, in: Proceedings of the Fifth International World Wide Web Conference, Paris, France, May 1996.

[6] S. Chakrabarti, B. Dom, D. Gibson, R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, Experiments in topic distillation, SIGIR Workshop on Hypertext IR, 1998.

[7] S. Chakrabarti, B. Dom, P. Raghavan, S. Rajagopalan, D. Gibson, J. Kleinberg, Automatic resource compilation by analyzing hyperlink structure and associated text, in: Proceedings of the Seventh World Wide Web Conference, Brisbane, Australia, April 1998.

[8] J. Cho, H. Garcia-Molina, L. Page, Efficient crawling through URL ordering, in: Proceedings of the Seventh World Wide Web Conference, Brisbane, Australia, April 1998.

[9] D. Cohn, H. Chang, Learning to probabilistically identify authorative documents, in: Proceedings of the 17th International Conference on Machine Learning, California, 2000.

[10] J. Dean, M. Henzinger, Finding related pages in the World Wide Web, in: Proceedings of the Eighth World Wide Web Conference, Toronto, Canada, May 1999.

[11] D. Dhyani, S.S. Bhowmick, W.K. Ng, Modeling and predicting Web page accesses using Burrell's method, in: Proceedings of the 3rd International Conference on Electronic Commerce and Web Technologies (EC-WEB 2002), France, 2002.

[12] D. Dhyani, S.S. Bhowmick, W.K. Ng, Web informetrics: extending classical informetrics to the Web, in: Proceedings of the 3rd International Workshop on Management Information on the Web (MIW 2002), France, 2002.

[13] D. Dhyani, S.S. Bhowmick, W.K. Ng, Deriving and verifying statistical distribution of a hyperlinked-based Web page quality metric, in: Proceedings of the 13th International Conference on Databases and Expert Systems Applications (DEXA 2002), France, 2002.

[14] L. Egghe, R. Rousseau, Introduction to Informetrics, Elsevier Science Publishers, 1990.

[15] D. Gibson, J. Kleinberg, P. Raghavan, Inferring Web communities from link topology, in: Proceedings of the Ninth ACM Conference on Hypertext and Hypermedia, Pittsburgh, PA, June 1998.

[16] D. Hawking, N. Craswell, P. Thislewaite, D. Harman, Results and challenges in Web search evaluation, in: Proceedings of the Eighth World Wide Web Conference, Toronto, Canada, May 1999.

[17] G. Hardin, The last canute, in: Scientific Monthly, 1946, George Allen and Unwin, London, 1965.

[18] M. Henzinger, A. Heydon, Measuring index quality using random walks on the Web, in: Proceedings of the Eighth World Wide Web Conference, Toronto, Canada, May 1999.

[19] T. Haveliwala, Efficient computation of PageRank, Stanford Technical Report, 1999.

[20] J. Kleinberg, Authoritative sources in a hyperlinked environment, in: Proceedings of the ACM-SIAM Symposium on Discrete Algorithms, 1998.

[21] J. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, The Web as a graph: measurements, models, and methods, in: Proceedings of the Fifth International Conference on Computing and Combinatorics, COCOON, Tokyo, Japan, July 1999.

[22] D. Lee, H. Chuang, K. Seamons, Effectiveness of document ranking and relevance feedback techniques, IEEE Software 14 (2) (1997) 67–75.

[23] R. Lempel, S. Moran, The stochastic approach for link structure analysis (SALSA) and the TKC effect, in: Proceedings of the Ninth World Wide Web Conference, 2000.

[24] R. Lempel, A. Soffer, PicASHOW: pictorial authority search by hyperlinks on the Web, in: Proceedings of the Tenth International World Wide Web Conference, Hong Kong, 2001.

[25] R. Miller, K. Bharat, SPHINX: a framework for creating personal, site-specific Web crawlers, in: Proceedings of the Seventh International World Wide Web Conference, Brisbane, Australia, April 1998.

[26] M. Marchiori, The quest for correct information on the Web: hyper search engines, in: Proceedings of the Sixth World Wide Web Conference, Santa Clara, California, April 1997.

[27] D. Rafiei, A. Mendelzon, What is this page known for? Computing Web page reputations, in: Proceedings of the Ninth World Wide Web Conference, Amsterdam, Netherlands, May 2000.

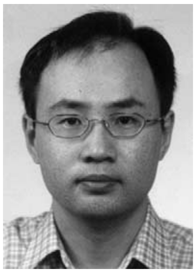
[28] J. Pitkow, C. Kehoe, Emerging trends in the www user population, Communications of the ACM 39 (6) (1996).

[29] J. Pitkow, In search of reliable usage data on the WWW, in: Proceedings of the Sixth World Wide Web Conference, Santa Clara, California, April 1997.

[30] B. Yuwono, S. Lam, J. Ying, D. Lee, A World Wide Web resource discovery system, in: Proceedings of the Fourth International World Wide Web Conference, 1995.

Devanshu Dhyani received his Master's degree in Computer Engineering from Nanyang Technological University, Singapore in 2000. He is currently working as consultant in Boston Consulting Group (BCG), Singapore.

Sourav S. Bhowmick received his Ph.D. in Computer Engineering from Nanyang Technological University, Singapore in 2001. He is an Assistant Professor of the School of Computer Engineering at the Nanyang Technological University. His current research interests include XML data management, mobile data management, biological data management, Web warehousing and Web mining. He has published more than 50 papers in major international database conferences and journals such as VLDB, ICDE, ICDCS, ER, IEEE TKDE, ACM CS, DKE and DEXA. He is serving as PC member of various database conferences and workshops and reviewer for various database journals. He is a member of the ACM and IEEE Computer Society.



Wee Keong Ng is an Assistant Professor of the School of Computer Engineering at the Nanyang Technological University, Singapore. He obtained his M.Sc. and Ph.D. degrees from the University of Michigan, Ann Arbor in 1994 and 1996 respectively. He works and publishes widely in the areas of Web warehousing, information extraction, electronic commerce and data mining. He has organized and chaired international workshops, including tutorials, and has actively served in the program committees of numerous international conferences. He is a member of the ACM and IEEE Computer Society.