Web Page Scoring Systems for Horizontal and Vertical Search

Michelangelo Diligenti
Dipartimento di Ingegneria dell'Informazione
Via Roma 56
Siena, Italy
diligmic@dii.unisi.it

Marco Gori
Dipartimento di Ingegneria dell'Informazione
Via Roma 56
Siena, Italy
marco@dii.unisi.it

Marco Maggini
Dipartimento di Ingegneria dell'Informazione
Via Roma 56
Siena, Italy
maggini@dii.unisi.it

ABSTRACT

Page ranking is a fundamental step towards the construction of effective search engines for both generic (*horizontal*) and focused (*vertical*) search. Ranking schemes for horizontal search like the PageRank algorithm used by Google operate on the topology of the graph, regardless of the page content. On the other hand, the recent development of vertical portals (*vortals*) makes it useful to adopt scoring systems focussed on the topic and taking the page content into account.

In this paper, we propose a general framework for Web Page Scoring Systems (WPSS) which incorporates and extends many of the relevant models proposed in the literature. Finally, experimental results are given to assess the features of the proposed scoring systems with special emphasis on vertical search.

Categories and Subject Descriptors

F.2.2 [**Analysis of Algorithms and Problem Complexity**]: Nonnumerical Algorithms and Problems—*Sorting and Searching*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Information Filtering*; H.5.4 [**Information Interfaces and Presentation**]: Hypertext/Hypermedia

General Terms

Algorithms

Keywords

Web Page Scoring Systems, Random Walks, HITS, PageRank, Focused PageRank

1. INTRODUCTION

The analysis of the hyperlinks on the Web [1] can significantly increase the capability of search engines. A simple counting of the number of references does not take into account the fact that not all the citations have the same authority. PageRank[1][2] is a noticeable example of a topological-based ranking criterion. An interesting example of query-dependent criteria is given in [3]. User queries are issued to a search engine in order to create a set of seed pages. Crawling the Web forward and backward from that seed is performed to mirror the Web portion containing the information which is likely to be useful. A ranking criterion based on topological analyses can be applied to the pages belonging to the selected Web portion. Very interesting results in this direction have been proposed in [4, 5, 6]. In [7] a Bayesian approach is used to compute hub and authorities, whereas in [8] both topological information and information about page content are included in distillation of information sources performed by a Bayesian approach.

Generally speaking, the ranking of hypertextual documents is expected to take into account the *reputation of the source*, the *page updating frequency*, the *popularity*, the *speed access*, the degree of authority, and the degree of hubness.

The page rank of hyper-textual documents can be thought of as a function of the document content and the hyperlinks. In this paper, we propose a general framework for Web Page Scoring Systems (WPSS) which incorporates and extends many of the relevant models proposed in the literature. The general web page scoring model proposed in this paper extends both PageRank [2] and the HITS scheme [3]. In addition, the proposed model exhibits a number of novel features, which turn out to be very useful especially for focused (*vertical*) search. The content of the pages is combined with the web graphical structure giving rise to scoring mechanisms which which are focused on a specific topic. Moreover, in the proposed model, vertical search schemes can take into account the mutual relationship amongst different topics. In so doing, the discovery of pages with high score for a given topic affects the score of pages with related topics.

Experimental results were carried out to assess the features of the proposed scoring systems with special emphasis on vertical search. The very promising experimental results reported in the paper provide a clear validation of the proposed general scheme for web page scoring systems.

2. PAGE RANK AND RANDOM WALKS

Random walk theory has been widely used to compute the absolute relevance of a page in the Web [2, 6]. The Web is represented as a graph $\boldsymbol{G}$, where each Web page is a node and a link between two nodes represents a hyperlink between the associated pages. A common assumption is that the relevance x_p of page p is represented by the probability of ending up in that page during a walk on this graph.

In the general framework we propose, we consider a complete model of the behavior of a user surfing the Web. We assume that a Web surfer can perform one out of four atomic actions at each step of his/her traversal of the Web graph:

- j jump to a node of the graph;

[1]http://www.google.com

WWW2002, May 7–11, 2002, Honolulu, Hawaii, USA.
ACM 1-58113-449-5/02/0005.

- l follow a hyperlink from the current page;
- b follow a back-link (a hyperlink in the inverse direction);
- s stay in the same node.

Thus the set of the atomic actions used to move on the Web is $\mathcal{O} = \{j, l, b, s\}$.

We assume that the behavior of the surfer depends on the page he is currently visiting. The action he/she will decide to take will depend on the page contents and the links it contains. For example if the current page is interesting to the surfer, it is likely he/she will follow a hyperlink contained in the page. Whereas, if the page is not interesting, the surfer will likely jump to a different page not linked by the current one. We can model the user behavior by a set of probabilities which depend on the current page:

- $x(l|q)$ the probability of following one hyperlink from page q,
- $x(b|q)$ the probability of following one back-link from page q,
- $x(j|q)$ the probability of jumping from page q,
- $x(s|q)$ the probability of remaining in page q.

These values must satisfy the normalization constraint

$$\sum_{o \in \mathcal{O}} x(o|q) = 1$$

Most of these actions need to specify their targets. Assuming that the surfer's behavior is time-invariant, then we can model the targets for jumps, hyperlink or back-link choices by using the following parameters:

- $x(p|q, j)$ the probability of jumping from page q to page p;
- $x(p|q, l)$ the probability of selecting a hyperlink from page q to page p; this value is not null only for the pages p linked directly by page q, i.e. $p \in ch(q)$, being $ch(q)$ the set of the children of node q in the graph $\boldsymbol{G}$;
- $x(p|q, b)$ the probability of going back from page q to page p; this value is not null only for the pages p which link directly page q, i.e. $p \in pa(q)$, being $pa(q)$ the set of the parents of node q in the graph $\boldsymbol{G}$.

These sets of values must satisfy the following probability normalization constraints for each page $q \in \boldsymbol{G}$: $\sum_{p \in \boldsymbol{G}} x(p|q, j) = 1$, $\sum_{p \in ch(q)} x(p|q, l) = 1$, $\sum_{p \in pa(q)} x(p|q, b) = 1$.

The model considers a temporal sequence of actions performed by the surfer and it can be used to compute the probability that the surfer is located in page p at time t, $x_p(t)$. The probability distribution on the pages of the Web is updated by taking into account the possible actions at time $t + 1$ using the following equation

$$x_p(t+1) = \sum_{q \in \boldsymbol{G}} x(p|q) \cdot x_q(t), \tag{1}$$

where the probability $x(p|q)$ of going from page q to page p is obtained by considering the action which can be performed by the surfer. Thus, using the previous definitions for the actions, the equation can be rewritten as

$$\begin{aligned} x_p(t+1) = &\sum_{q \in \boldsymbol{G}} x(p|q, j) \cdot x(j|q) \cdot x_q(t) + \\ &\sum_{q \in pa(p)} x(p|q, l) \cdot x(l|q) \cdot x_q(t) + \\ + &\sum_{q \in ch(p)} x(p|q, b) \cdot x(b|q) \cdot x_q(t) + x(s|p) \cdot x_p(t) \end{aligned} \tag{2}$$

These probabilities can be collected in a N-dimensional vector $\boldsymbol{x}(t)$, being N the number of pages in the Web graph $\boldsymbol{G}$, and the probability update equations can be rewritten in a matrix form. The probabilities of moving from a page q to a page p given an action can be organized into the following matrices:

- the *forward* matrix $\boldsymbol{\Delta}$ whose element (p, q) is the probability $x(p|q, l)$;
- the *backward* matrix $\boldsymbol{\Gamma}$ collecting the probabilities $x(p|q, b)$;
- the *jump* matrix $\boldsymbol{\Sigma}$ which is defined by the jump probabilities $x(p|q, j)$.

The forward and backward matrices are related to the Web adjacency matrix $\boldsymbol{W}$ whose entries are 1 if page p links page q. In particular, the forward matrix $\boldsymbol{\Delta}$ has non null entries only in the positions corresponding to 1s in matrix $\boldsymbol{W}$, and the backward matrix $\boldsymbol{\Gamma}$ has non null entries in the positions corresponding to 1s in $\boldsymbol{W}'$.

Further, we can define the set of *action* matrices which collect the probabilities of taking one of the possible actions from a given page q. These are $N \times N$ diagonal matrices defined as follows: $\boldsymbol{D}_j$ whose diagonal values (q, q) are the probabilities $x(j|q)$, $\boldsymbol{D}_l$ collecting the probabilities $x(l|q)$, $\boldsymbol{D}_b$ containing the values $x(b|q)$, and $\boldsymbol{D}_s$ having on the diagonal the probabilities $x(s|q)$.

Hence, equation (2) can be written in matrix form as

$$\begin{aligned} \boldsymbol{x}(t+1) = &(\boldsymbol{\Sigma} \cdot \boldsymbol{D}_j)' \boldsymbol{x}(t) + (\boldsymbol{\Delta} \cdot \boldsymbol{D}_l)' \boldsymbol{x}(t) + \\ &(\boldsymbol{\Gamma} \cdot \boldsymbol{D}_b)' \boldsymbol{x}(t) + (\boldsymbol{D}_s)' \boldsymbol{x}(t) \end{aligned} \tag{3}$$

The transition matrix $\boldsymbol{T}$ used to update the probability distribution is

$$\boldsymbol{T} = (\boldsymbol{\Sigma} \cdot \boldsymbol{D}_j + \boldsymbol{\Delta} \cdot \boldsymbol{D}_l + \boldsymbol{\Gamma} \cdot \boldsymbol{D}_b + \boldsymbol{D}_s)' \tag{4}$$

Using this definition, the equation (3) can be written as

$$\boldsymbol{x}(t+1) = \boldsymbol{T} \cdot \boldsymbol{x}(t) \tag{5}$$

Starting from a given initial distribution $\boldsymbol{x}(0)$ equation (5) can be applied recursively to compute the probability distribution at a given time step t yielding

$$\boldsymbol{x}(t) = \boldsymbol{T}^t \cdot \boldsymbol{x}(0). \tag{6}$$

In order to define an absolute page rank for the pages on the Web, we consider the stationary distribution of the Markov chain defined by the previous equations. $\boldsymbol{T}'$ is the state transition matrix of the Markov chain. $\boldsymbol{T}'$ is stable, since $\boldsymbol{T}'$ is a stochastic matrix having its maximum eigenvalue equal to 1. Since the state vector $\boldsymbol{x}(t)$ evolves following the equation of a Markov Chain, it is guaranteed that if $\sum_{q \in \boldsymbol{G}} \boldsymbol{x}_q(0) = 1$, then $\sum_{q \in \boldsymbol{G}} \boldsymbol{x}_q(t) = 1,\ t = 1, 2, \ldots$.

By applying the results on Markov chains (see e.g. [9]), we can prove the following proposition.

PROPOSITION 2.1. *If* $x(j|q) \neq 0 \wedge x(p|q, j) \neq 0,\ \forall p, q :\in \boldsymbol{G}$ *then* $lim_{t \to \infty} \boldsymbol{x}(t) = \boldsymbol{x}^\star$ *where* $\boldsymbol{x}^\star$ *does not depend on the initial state vector* $\boldsymbol{x}(0)$.

Proof: Because of the hypotheses, $\boldsymbol{\Sigma} \cdot \boldsymbol{D}_j$ is strictly positive, i.e. all its entries are greater then 0. Since the transition matrix $\boldsymbol{T}$ of the Markov chain is obtained by adding non-negative matrices, then also the transition matrix $\boldsymbol{T}$ is strictly positive. Thus, the resulting Markov Chain is *irreducible* and consequently it has a unique stationary distribution given by the solution of the equation $\boldsymbol{x}^\star = \boldsymbol{T}\boldsymbol{x}^\star$, where $\boldsymbol{x}^\star$ satisfies $(\boldsymbol{x}^\star)'\mathbf{\Pi}$ being $\mathbf{\Pi}$ the N-dimensional vector whose entries are all 1s. □

2.1 Uniform Jump Probabilities

In order to simplify the general model proposed so far, we can introduce some assumptions on the probability matrices. A possible choice is to consider some actions to be independent on the current page. A first hypothesis we investigate is the case of jump probabilities which are independent on the starting page q. This choice models a surfer who decides to make random jumps from a given page to another page with uniform probability. Thus, the jump matrix $\boldsymbol{\Sigma}$ has all the entries equal to $x(p|q,j) = x(p|j) = \frac{1}{N}$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \frac{1}{N} & \frac{1}{N} & \cdots & \frac{1}{N} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{N} & \frac{1}{N} & \cdots & \frac{1}{N} \end{bmatrix}$$

Moreover, we suppose that also the probability of choosing a jump among the available actions does not depend on the page, i.e. $x(j|p) = d_j, \forall p : p \in \boldsymbol{G}$. Under these two assumptions, equation (3) becomes

$$\boldsymbol{x}(t+1) = \frac{d_j}{N}\mathbf{\Pi} + (\boldsymbol{\Delta} \cdot \boldsymbol{D}_l + \boldsymbol{\Gamma} \cdot \boldsymbol{D}_b + \boldsymbol{D}_s)'\boldsymbol{x}(t) \qquad (7)$$

because $(\boldsymbol{\Sigma} \cdot \boldsymbol{D}_j)' \cdot \boldsymbol{x}(t) = (d_j/N) \cdot \mathbf{\Pi}$ as $\sum_{p \in \boldsymbol{G}} \boldsymbol{x}_p(t) = 1$.

If we define $\boldsymbol{R} = (\boldsymbol{\Delta} \cdot \boldsymbol{D}_l + \boldsymbol{\Gamma} \cdot \boldsymbol{D}_b + \boldsymbol{D}_s)'$, the solution of equation (7) is

$$\boldsymbol{x}(t) = \boldsymbol{R}^t \cdot \boldsymbol{x}(0) + \frac{d_j}{N}\left(\sum_{j=0}^{t} \boldsymbol{R}^{t-1-j}\right) \cdot \mathbf{\Pi} \qquad (8)$$

Using the Frobenius theorem on matrices, the following bound on the maximum eigenvalue of $\boldsymbol{R}$ can be derived

$$\begin{aligned} \lambda_{max}(\boldsymbol{R}) &= \lambda_{max}(\boldsymbol{R}') = \lambda_{max}(\boldsymbol{\Delta} \cdot \boldsymbol{D}_l) + \\ &\quad + \lambda_{max}(\boldsymbol{\Gamma} \cdot \boldsymbol{D}_b) + \lambda_{max}(\boldsymbol{D}_s) \\ &\leq \max_{q=1,\ldots,N} (x(l|q) \cdot \lambda_{max}(\boldsymbol{\Delta}) + \\ &\quad + x(b|q) \cdot \lambda_{max}(\boldsymbol{\Gamma}) + x(s|q)) = \\ &= \max_{q=1,\ldots,N} (x(l|q) + x(b|q) + x(s|q)) < \\ &< \max_{q=1,\ldots,N} (x(l|q) + x(b|q) + x(j|q) + x(s|q)) \\ &= 1 \end{aligned} \qquad (9)$$

because $\lambda_{max}(\boldsymbol{\Delta}) = \lambda_{max}(\boldsymbol{\Gamma}) = 1$. Hence, in equation (8) the first term vanishes as $t \to \infty$, i.e.

$$\lim_{t \to \infty} \boldsymbol{R}^t = \boldsymbol{O} \qquad (10)$$

where $\boldsymbol{O}$ is the matrix having all elements equal to 0. Thus, when $t \to \infty$, the probability distribution $\boldsymbol{x}(t)$ converges to

$$\boldsymbol{x}^\star = \lim_{t \to \infty} \boldsymbol{x}(t) = \frac{d_j}{N}\left(\sum_{t=0}^{\infty} \boldsymbol{R}^t\right) \cdot \mathbf{\Pi} \qquad (11)$$

Since $\lambda_{max}(\boldsymbol{R}) < 1$ as shown in equation (9), it can be proven that the previous equation converges to

$$\boldsymbol{x}^\star = \frac{d_j}{N}(\boldsymbol{I} - \boldsymbol{R})^{-1} \cdot \mathbf{\Pi} \qquad (12)$$

As it is stated by proposition 2.1, the value of $\boldsymbol{x}^\star$ does not depend on the choice of the initial state vector $\boldsymbol{x}(0)$.

2.2 Multiple State Model

A model based on a single variable may not capture the complex relationships among Web pages when trying to model their importance. Ranking schemes based on multiple variables have been proposed in [3, 6], where a pair of variables are used to represents the concepts of *hubness* and *authority* of a page. In the probabilistic framework described so far, we can define a multivariable scheme by considering a pool of Web surfers each described by a single variable. Each surfer is characterized by his/her own way of browsing the Web modeled by using different parameter values in each state transition equation. By choosing proper values for these parameters we can choose different policies in evaluating the absolute importance of the pages. Moreover, the surfers may interact by accepting *suggestions* from each other.

In order to model the activity of M different surfers, we use a set of state variables $x_q^{(i)}(t)$ which represent the probability of each surfer i to be visiting page q at time t. The interaction among the surfers is modeled by a set of parameters which define the probability of the surfer k to accept the suggestion of the surfer i, thus jumping from the page he/she is visiting to the one visited by the surfer i. This interaction happens before the choice of the actions described previously. If we hypothesize that the interaction does not depend on the page the surfer k is currently visiting, the degree of interaction with the surfer i is modeled by the value $b(i|k)$ which represents the probability for the surfer k of jumping to the page visited by the surfer i. These values must satisfy the probability normalization constraint $\sum_{s=1}^{M} b(s|k) = 1$.

As an example, suppose that there are two surfers, "the novice" and "the expert". Surfer s_1, the novice, blindly trusts the suggestions of surfer s_2 as he/she believes s_2 is an expert in discovering authoritative pages, whereas s_1 does not trust at all his/her own capabilities. In this case the complete dependence of the novice on the expert is modeled by choosing $b(1|1) = 0$ and $b(2|1) = 1$, i.e. the novice chooses the page visited by the expert with probability equal to 1.

Before taking any action, the surfer i repositions himself/herself in page p with probability $v_p^{(i)}(t)$ looking at the suggestions of the other surfers. This probability is computed as

$$v_p^{(i)}(t) = \sum_{s=1}^{M} b(s|i) x_p^{(s)}(t) \qquad (13)$$

Thus, when computing the new probability distribution $x_p^{(i)}(t+1)$ due to the action taken at time t by the surfer i, we consider the distribution $v_p^{(i)}(t)$ instead of $x_p^{(i)}(t)$ when applying equation (2). Thus, the transition function is defined as follows

$$\begin{aligned} x_p^{(i)}(t+1) &= \sum_{q \in \boldsymbol{G}} x^{(i)}(p|q,j) \cdot x^{(i)}(j|q) \cdot \sum_{s=1}^{M} b(s|i) x_q^s(t) + \\ &+ \sum_{q \in pa(p)} x^{(i)}(p|q,l) \cdot x^{(i)}(l|q) \cdot \sum_{s=1}^{M} b(s|i) x_q^s(t) + \\ &+ \sum_{q \in ch(p)} x^{(i)}(p|q,b) \cdot x^{(i)}(b|q) \cdot \sum_{s=1}^{M} b(s|i) x_q^s(t) + \end{aligned}$$

$$+x^{(i)}(s|p) \cdot \sum_{s=1}^{M} b(s|i) x_p^s(t) \qquad i = 0, \ldots, M \tag{14}$$

When considering M surfers, the score vectors of each surfer $\boldsymbol{x}^{(i)}(t)$ can be collected as the columns of a matrix $\boldsymbol{X}(t)$. Moreover, we define the $[M \times M]$ matrix $\boldsymbol{A}$ which collects the values $b(i|k)$. The matrix $\boldsymbol{A}$ will be referred to as the *interaction* matrix. Finally, each surfer i will be described by his/her own forward, backward and jump matrices $\boldsymbol{\Sigma}^{(i)}$, $\boldsymbol{\Delta}^{(i)}$, $\boldsymbol{\Gamma}^{(i)}$ and by the action matrices $\boldsymbol{D}_j^{(i)}$, $\boldsymbol{D}_l^{(i)}$, $\boldsymbol{D}_b^{(i)}$, $\boldsymbol{D}_s^{(i)}$. Thus, the transition matrix for the Markov chain associated to the surfer i is $\boldsymbol{T}^{(i)} = \left(\boldsymbol{\Sigma}^{(i)} \cdot \boldsymbol{D}_j^{(i)} + \boldsymbol{\Delta}^{(i)} \cdot \boldsymbol{D}_l^{(i)} + \boldsymbol{\Gamma} \cdot \boldsymbol{D}_b^{(i)} + \boldsymbol{D}_s^{(i)}\right)'$.

Using these definitions, the set of M interacting surfers can be described by rewriting equation (14) as a set of matrix equations as follows

$$\begin{cases} \boldsymbol{x}^{(1)}(t+1) = \boldsymbol{T}^{(1)} \cdot \boldsymbol{X}(t) \cdot \boldsymbol{A}^{(1)} \\ \vdots \\ \boldsymbol{x}^{(M)}(t+1) = \boldsymbol{T}^{(M)} \cdot \boldsymbol{X}(t) \cdot \boldsymbol{A}^{(M)} \end{cases} \tag{15}$$

When the surfers are independent on each other (i.e. $b(i|i) = 1$, $i = 0, \ldots, M$, and $b(i|j) = 0$, $i = 0, \ldots, M$, $j = 0, \ldots, M$, $j \neq i$), the model reduces to M models as described by equation (6).

3. HORIZONTAL WPSS

Horizontal WPSSs do not consider any information on the page contents and produce the rank vector using just the topological characteristics of the Web graph. In this section we show how these scoring systems can be described in the proposed probabilistic framework. In particular we derive the two most popular Page Scoring Systems, *PageRank* and HITS, as special cases.

3.1 PageRank

The Google search engine employs a ranking scheme based on a random walk model defined by a single state variable. Only two actions are considered: the surfer jumps to a new random page with probability $1 - d$ or he/she follows one link from the current page with probability d. The Google ranking scheme, called *PageRank*, can be described in the general probabilistic framework of equation (2), by choosing its parameters as follows. First, the probabilities of following a back-link $x(b|p)$ and of remaining in any page $x(s|p)$ are null for all the pages p. Then, as stated above, the probability of performing a random jump is $x(j|p) = 1 - d$ for any page p, whereas the probability of following a hyperlink contained in the page p is also a constant, i.e. $x(l|p) = d$. Given that a jump is taken, its target is selected using a uniform probability distribution over all the N Web pages, i.e. $x(p|j) = 1/N$, $\forall p \in \boldsymbol{G}$. Finally, the probability of following the hyperlink from q to p does not depend on the page p, i.e. $x(p|q, l) = \alpha_q$. In order to meet the normalization constraint, $\alpha_q = 1/h_q$ where h_q is the number of links exiting from page q (the page hubness). This assumption makes the surfer *random*, since we define a uniform probability distribution among all the outgoing links.

This last hypothesis cannot be met by pages which do not contain any links to other pages. A page with no out-links is called *sink* page, since it would behave just like a score sink in the PageRank propagation scheme. In order to keep the probabilistic interpretation of PageRank, all sink nodes must be removed. The page rank of sinks is then computed from the page ranks of their parents.

Under all these assumptions equation (2) can be rewritten as

$$x_p(t+1) = \frac{(1-d)}{N} \cdot \sum_{q \in \boldsymbol{G}} x_q(t) + d \sum_{q \in pa(p)} \alpha_q \cdot x_q(t) \tag{16}$$

Since the probabilistic interpretation is valid, it holds that

$$\sum_{q \in \boldsymbol{G}} x_q(t) = 1$$

and, then, equation (16) becomes

$$x_p(t+1) = \frac{(1-d)}{N} + d \sum_{q \in pa(p)} \alpha_q \cdot x_q(t) \tag{17}$$

Since $0 < d < 1$, $x(j|p) = 1 - d > 0$ for each page and then equation (12) guarantees that the Google's PageRank converges to a distribution of page scores that does not depend on the initial distribution.

In order to apply the Google's PageRank scheme without removing the sink nodes, we can introduce the following modification to the original equations. Since no links can be followed from a sink node q, $x(l|q)$ must be equal to 0 and $x(j|q)$ equal to 1. Thus, when there are sinks, $x(j|q)$ is defined as

$$\begin{cases} x(j|q) = 1 - d & \textbf{if } ch(q) \neq \emptyset \\ x(j|q) = 1 & \textbf{if } ch(q) = \emptyset \end{cases} \tag{18}$$

In this case the contribution of the jump probabilities does not sum to a constant term as it happens in equation (17), but the value $x(p|j,t) = \frac{1}{N} \sum_{q \in \boldsymbol{G}} x(j|q) x_q(t)$ must be computed at the beginning of each iteration. This is the computational scheme we used in our experiments.

3.2 The HITS Ranking System

The *HITS* algorithm was proposed to model authoritative documents only relying on the information hidden in the connections among them due to co-citation or web hyperlinks [3]. In this formulation the Web pages are divided into two page classes: pages which are information sources (*authorities*) and pages which link to information sources (*hubs*). The HITS algorithm assigns two numbers to each page p, the page *authority* and the page *hubness* in order to model the importance of the page. These values are computed by applying iteratively the following equations

$$\begin{cases} a_q(t+1) = \sum_{p \in pa(q)_i} h_p(t) \\ h_q(t+1) = \sum_{p \in ch(q)} a_p(t), \end{cases} \tag{19}$$

where a_q indicates the authority of page q and h_q its hubness. If $\boldsymbol{a}(t)$ is the vector of the authorities at step t, and $\boldsymbol{h}(t)$ is the hubness vector at step t, the previous equation can be rewritten in matrix form as

$$\begin{cases} \boldsymbol{a}(t+1) = \boldsymbol{W} \cdot \boldsymbol{h}(t) \\ \boldsymbol{h}(t+1) = \boldsymbol{W}' \cdot \boldsymbol{a}(t), \end{cases} \tag{20}$$

where $\boldsymbol{W}$ is the adjacency matrix of the Web graph. It is trivial to demonstrate that as t tends to infinity, the direction of authority vector tends to be parallel to the main eigenvector of the $\boldsymbol{W} \cdot \boldsymbol{W}'$ matrix, whereas the hubness vector tends to be parallel to the main eigenvector of the $\boldsymbol{W}' \cdot \boldsymbol{W}$ matrix.

The HITS ranking scheme can be represented in the general Web surfer framework, even if some of the assumptions will violate the probabilistic interpretation. Since HITS uses two state variables, the hubness and the authority of a page, the corresponding random walk model is a multiple state scheme based on the activity of two surfers. Surfer 1 is associated to the hubness of pages whereas

surfer 2 is associated to the authority of pages. For both surfers the probabilities of remaining in the same page $x^{(i)}(s|p)$ and of jumping to a random page $x^{(i)}(s|p)$ are null. Surfer 1 never follows a link, i.e. $x^{(1)}(l|p) = 0,\ \forall p \in \boldsymbol{G}$ whereas he/she always follows a back-link, i.e. $x^{(1)}(b|p) = 1,\ \forall p \in \boldsymbol{G}$. Because of this, the HITS computation violates the probability normalization constraints, since $\sum_{p \in pa(q)} x^{(1)}(p|q,b) = |pa(q)| \geq 1$.

Surfer 2 has the opposite behavior with respect to surfer 1. He/she always follows a link, i.e. $x^{(2)}(l|p) = 1,\ \forall p \in \boldsymbol{G}$, and he/she never follows a back-link, i.e. $x^{(2)}(b|p) = 0$. In this case the normalization constraint is violated for the values of $x^{(2)}(l|p)$ because $\sum_{p \in ch(q)} x^{(1)}(p|q,b) = |ch(q)| \geq 1$.

Under these assumptions $\boldsymbol{D}_b^{(1)} = \boldsymbol{I},\ \boldsymbol{D}_l^{(2)} = \boldsymbol{I}$, being I the identity matrix, whereas $\boldsymbol{D}_j^{(1)}$, $\boldsymbol{D}_l^{(1)}$, $\boldsymbol{D}_s^{(1)}$, $\boldsymbol{D}_j^{(2)}$, $\boldsymbol{D}_b^{(2)}$, $\boldsymbol{D}_s^{(2)}$ are all equal to the null matrix.

The interaction between the surfers is described by the matrix:

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \tag{21}$$

The interpretation of the interactions represented by this matrix is that surfer 1 considers surfer 2 as an expert in discovering authorities and always moves to the position suggested by that surfer before acting. On the other hand, surfer 2 considers surfer 1 as an expert in finding hubs and then he/she always moves to the position suggested by that surfer before choosing the next action.

In this case equation (15) is

$$\begin{cases} \boldsymbol{x}^{(1)}(t+1) = \left(\boldsymbol{\Delta}^{(1)}\right)' \cdot \boldsymbol{X}(t) \cdot \boldsymbol{A}^{(1)} \\ \boldsymbol{x}^{(2)}(t+1) = \left(\boldsymbol{\Gamma}^{(2)}\right)' \cdot \boldsymbol{X}(t) \cdot \boldsymbol{A}^{(2)} \end{cases} \tag{22}$$

Using equation (21) and the HITS assumption $\boldsymbol{\Delta}^{(1)\prime} = \boldsymbol{W}'$, $\boldsymbol{\Gamma}^{(2)\prime} = \boldsymbol{W}$ we obtain

$$\begin{cases} \boldsymbol{x}^{(1)}(t+1) = \boldsymbol{W}' \cdot \boldsymbol{x}^{(2)}(t) \\ \boldsymbol{x}^{(2)}(t+1) = \boldsymbol{W} \cdot \boldsymbol{x}^{(1)}(t) \end{cases} \tag{23}$$

which, redefining $\boldsymbol{a}(t) = \boldsymbol{x}^{(1)}(t)$ and $\boldsymbol{h}(t) = \boldsymbol{x}^{(2)}(t)$, is equivalent to the HITS computation of equation (20).

The HITS model violates the probabilistic interpretation and this makes the computation unstable, since the $\boldsymbol{W} \cdot \boldsymbol{W}'$ matrix has a principal eigenvalue much larger then 1. Hence, unlike Google's PageRank, the HITS algorithm needs the score to be normalized at the end of each iteration.

Finally, the HITS scheme suffers from other drawbacks. In particular, large highly connected communities of Web pages tend to attract the principal eigenvector of $\boldsymbol{W} \cdot \boldsymbol{W}'$, thus pushing to zero the relevance of all other pages. As a result the page scores tend to decrease rapidly to zero for pages outside those communities. Recently some heuristics have been proposed to avoid this problem even if such behavior can not be generally avoided because of the properties of the dynamic system associated to the HITS algorithm [10].

3.3 The PageRank-HITS model

PageRank is stable, it has a well defined behavior because of its probabilistic interpretation and it can be applied to large page collections without canceling the influence of the smallest Web communities. On the other hand, PageRank is sometimes too simple to take into account the complex relationships of Web page citations. HITS is not stable, only the largest Web community influences the ranking, and this does not allow the application of HITS to large page collections. On the other hand the hub and authority model can capture more than PageRank the relationships among Web pages. In this section we show that the proposed probabilistic framework allows to include the advantages of both approaches. We employ two surfers, each one implementing a bidirectional PageRank surfer. We assume that surfer 1 either follows a back-link with probability $x^{(1)}(b|p) = d^{(1)}$ or jumps to a random page with probability $x^{(1)}(j|p) = 1 - d^{(1)},\ \forall p \in \boldsymbol{G}$. Whereas surfer 2 either follows a forward link with probability $x^{(2)}(l|p) = d^{(2)}$ or jumps to a random page with probability $x^{(2)}(j|p) = 1 - d^{(2)},\ \forall p \in \boldsymbol{G}$.

Like in HITS, the interaction between the surfers is described by the matrix

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

In this case, equation (15) becomes

$$\begin{cases} \boldsymbol{x}^{(1)}(t+1) = \frac{(1-d^{(1)})}{N}\boldsymbol{\Pi} + d^{(1)}(\boldsymbol{\Gamma}^{(1)})' \cdot \boldsymbol{x}^{(2)}(t) \\ \boldsymbol{x}^{(2)}(t+1) = \frac{(1-d^{(2)})}{N}\boldsymbol{\Pi} + d^{(2)}(\boldsymbol{\Delta}^{(2)})' \cdot \boldsymbol{x}^{(1)}(t) \end{cases} \tag{24}$$

Further, we assume the independence of parameters $x^{(1)}(p|q,b)$ and $x^{(2)}(p|q,l)$ on the page p. Hence, it holds that $\boldsymbol{\Delta}^{(1)\prime} = \boldsymbol{W}' \cdot \boldsymbol{\Theta}$, $\boldsymbol{\Gamma}^{(2)\prime} = \boldsymbol{W} \cdot \boldsymbol{\Omega}$, where $\boldsymbol{\Omega}$ is the diagonal matrix with element (p,p) equal to $1/pa(p)$ and $\boldsymbol{\Theta}$ is the diagonal matrix with element (p,p) equal to $1/ch(p)$. Then:

$$\begin{cases} \boldsymbol{x}^{(1)}(t+1) = \frac{(1-d^{(1)})}{N}\boldsymbol{\Pi} + d^{(1)}\boldsymbol{W} \cdot \boldsymbol{\Omega} \cdot \boldsymbol{x}^{(2)}(t) \\ \boldsymbol{x}^{(2)}(t+1) = \frac{(1-d^{(2)})}{N}\boldsymbol{\Pi} + d^{(2)}\boldsymbol{W}' \cdot \boldsymbol{\Theta} \cdot \boldsymbol{x}^{(1)}(t) \end{cases} \tag{25}$$

This page rank is stable, the scores sum up to 1 and no normalization is required at the end of each iteration. Moreover, the two state variables can capture and process more complex relationships among pages.

In particular, setting $d^{(1)} = d^{(2)} = 1$ yields a normalized version of HITS, which has been proposed in [4].

4. VERTICAL WPSS

Horizontal WPSSs exploit the information provided by the Web graph topology. Different properties of the graph are evidenced by each model. For example the intuitive idea that a highly linked page is an absolute authority can be captured by PageRank or HITS schemes. However, when applying scoring techniques for focused search the page contents should be taken into account beside the graph topology. Vertical WPSSs aim at computing a relative ranking of pages when focusing on a specific topic. A vertical WPSS relies on the representation of the page content with a set of features (e.g. a set of keywords) and on a classifier which is used to assess the degree of relevance of the page with respect to the topic of interest. Basically the general probabilistic framework of WPSSs proposed in this paper can be used to define a vertical approach to page scoring. Several models can be derived which combine the ideas underlining the topology-based scoring and the topic relevance measure provided by text classifiers. In particular a text classifier can be used to compute proper values for the probabilities needed by the random walk model. As it is shown by the experimental results, vertical WPSSs can produce more accurate results in ranking topic specific pages.

4.1 Focused PageRank

In the PageRank framework, when choosing to follow a link in a page q each link has the same probability $1/ch(q)$ to be followed. Instead of the *random* surfer model, in the focused domain we can

consider the more realistic case of a surfer who follows the links according to the suggestions provided by a page classifier.

If the surfer is located at page q and the pages linked by page q have scores $s(ch_1(q)), \ldots, s(ch_{h_q}(q))$ by a topic classifier, the probability of the surfer to follow the i-th link is defined as

$$x(ch_i(q)|q,l) = \frac{s(ch_i(q))}{\sum_{j=0}^{h_q} s(ch_j(q))}. \quad (26)$$

Thus the forward matrix $\mathbf{\Delta}$ will depend on the classifier outputs on the target pages.

Hence, the modified equation to compute the combined page scores using a PageRank-like scheme is

$$x_p(t+1) = \frac{(1-d)}{N} + d \sum_{q \in pa(p)} x(p|q,l) \cdot x_q(t), \quad (27)$$

where $x(p|q,l)$ is computed as in equation (26).

This scoring system removes the assumption of complete randomness of the underlying Web surfer. In this case, the surfer is aware of what he/she is searching, and he/she will trust the classifier suggestions following the links with a probability proportional to the score of the page the links leads to. This allows us to derive a topic-specific page rank. For example: the "Microsoft" home is highly authoritative according to the topic-generic PageRank, whereas it is not highly authoritative when searching for "Perl" language tutorials, since even if that page gets many citations, most of these citations will be scarcely related to the target topic and then not significantly considered in the computation.

4.2 Double Focused PageRank

The focused PageRank model described previously uses a topic specific distribution for selecting the link to follow but the decision on the action to take does not depend on the contents of the current page. A more realistic model should take into account the fact that the decision about the action is usually dependent on the contents of the current page. For example, let us suppose that two surfers are searching for a "Perl Language tutorial", and that the first one is located at the page "http://www.perl.com", while the second is located at the page "http://www.cnn.com". Clearly it is more likely that the first surfer will decide to follow a link from the current page while the second one will prefer to jump to another page which is related to the topic he is interested in.

We can model this behavior by adapting the action probabilities using the contents of the current page, thus modeling a focused choice of the surfer's actions. In particular, the probability of following a hyperlink can be chosen to be proportional to the degree of relevance $s(q)$ of the current page with respect to the target, i.e.

$$x(l|p) = d_1 \cdot \frac{s(p)}{\max_{q \in \boldsymbol{G}} s(q)} \quad (28)$$

where $s(q)$ is computed by a text classifier.

On the other hand, the probability of jumping away from a page decreases proportionally to $s(q)$, i.e.

$$x(j|p) = 1 - d_1 \cdot \frac{s(p)}{\max_{q \in \boldsymbol{G}} s(q)} - d_2 \quad (29)$$

Finally, we assume that the probability of landing into page after a jump is proportional to its relevance $s(p)$, i.e.

$$x(p|j) = \frac{s(p)}{\sum_{q \in \boldsymbol{G}} s(q)} \quad (30)$$

Such modifications can be integrated into the focused PageRank proposed in section 4.1 to model a focused navigation more accurately. Equation (12) guarantees that the resulting scoring system is stable and that it converges to a score distribution independent from the initial distribution.

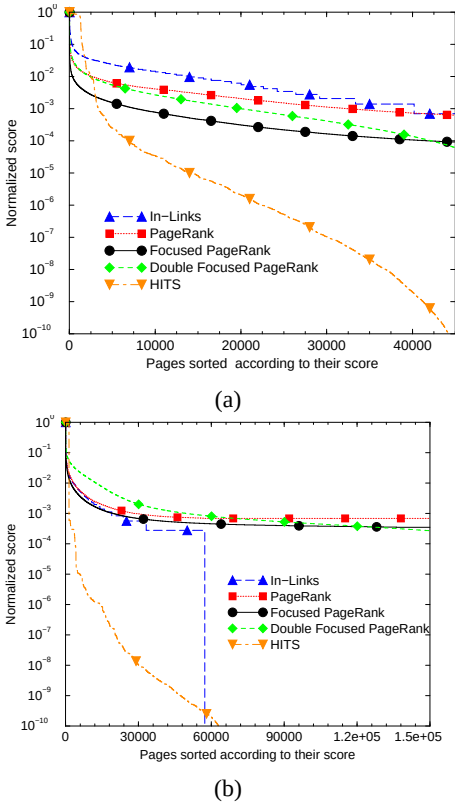


Figure 1: (a) The distribution of page scores for the topic "Linux". (b) The distribution of scores for the topic "cooking recipes". In both cases PageRank is much smoother that the HITS one. The focused versions of the PageRank are still smooth but concentrate the scores on a smaller set of authoritative pages which are more specific for the considered topic.

5. EXPERIMENTAL RESULTS

Using the focus crawler described in [11], we have performed two focus crawling sessions, downloading 150.000 pages for each single crawl. During the first session the crawler spidered the Web searching for pages on the topic "Linux". During the second session the crawler gathered pages on "cooking recipes". Each downloaded page was classified to assess its relevance with respect to the specific topic. Considering the hyperlinks contained in each page, two Web subgraphs were created to perform the evaluation of the different WPSSs proposed in the previous sections. For the second crawling session, the connectivity map of pages was pruned removing all links from a page to pages in the same site in order to reduce the "nepotism" of Web pages.

The topological structure of the graphs and the scores of the text classifiers were used to evaluate the following WPSSs:

- the "In-link" surfer. Such surfer is located in pages p with probability $x_p = s(p)/\sum_{q \in G} s(q)$ where $s(q)$ is the relevance of the page computed by the text classifier;
- the PageRank surfer;
- the Focused PageRank scheme as described in section 4.1;

PageRank
www.zdnet.com
www.google.com
search.internet.com/power_search
www.ibm.com
www.yahoo.com
www.ibm.com/planetwide/select
java.sun.com
www.osdn.com

HITS
www.openbsdapps.com/?page=category&categor...
www.openbsdapps.com/?page=category&categor...
www.openbsdapps.com/?page=category&categor...
www.openbsdapps.com/?page=category&categor...
www.openbsdapps.com/?page=category&categor...
www.openbsdapps.com/?page=category&categor...
www.openbsdapps.com/?page=newupdated&dat...
www.openbsdapps.com/?page=linkus

Figure 2: We report the 8 top score pages from a portion of the Web focused on the topic "Linux", using either the PageRank surfer, or a HITS surfer pool. For the HITS surfer pool we report the pages with the top authority value.

Focused PageRank
www.internet.com/sections/linux.html
www.slackware.com
www.linux.org
www.zdnet.com
jobs.osdn.com
www.yahoo.com
www.linux.org/books/index.html
www.python.org

Double Focused PageRank
www.internet.com/sections/linux.html
www.slackware.com
www.li.org
www.linux.org
www.linuxhq.com
www.slackware.org
www.linux.org/index.html
www.linuxusers.org

Figure 3: We report the 8 top score pages from a portion of the Web focused on the topic "Linux", using the proposed focused versions of the PageRank surfer.

- the Double Focused PageRank scheme described in section 4.2;
- the HITS surfer pool;
- the PageRank-HITS surfer pool.

5.1 The Distribution of Score Among Pages

We have performed an analysis of the distribution of page scores using the different algorithms proposed in this paper. For all the PageRank surfers (focused or not) we set the d parameter equal to 0.85. For each ranking function, we normalized the rank using its maximum value. We sorted the pages according to their ranks, then we plotted the distribution of the normalized rank values. Figure 1 reports the plots for the two categories, "Linux" and "cooking recipes". In both cases the HITS surfer assigns a score value significantly greater than zero only to the small set of pages associated to the main eigenvector of the connectivity matrix of the analyzed portion of the Web. On the other hand the PageRank surfer is more stable and its score distribution curve is smooth. This is the effect of the homogeneous term $1 - d$ in equation (17) and of the stability in the computation provided by the probabilistic interpretation. The Focused PageRank surfer and the Double Focused one still provide a smooth distribution. However, the focused page ranks are more concentrated around the origin. This reflects the fact that the vertical WPSSs are able to discriminate the authorities on the specific topic, whereas the classical PageRank scheme considers the authoritative pages regardless their topic.

5.2 Some Qualitative Results

Figures 2 and 3 show, respectively, the 8 top score pages for four different WPSSs on the database with pages on topic "Linux", while figures 4 and 5 reports the same results for the topic "cooking recipes".

As shown in figure 2, all pages distilled by the HITS algorithm come from the same site. In fact, a site which has many internal connections may acquire the principal eigenvector of the connectivity matrix associated to the considered portion of the Web graph, conquering all the top positions in the page rank and hiding all other resources.

Even the elimination of intra-site links does not improve the performances of the HITS algorithm. For example, as shown in the HITS section of figure 4, the Web site "www.allrecipe.com", which is subdivided into a collection of Web sites ("www.seafoodrecipes-.com", "www.cookierecipes.com", etc.) strongly connected among them, occupies all the top positions in the ranking list, hiding all other resources. In [10] the content of pages is considered in order to propagate relevance scores only over the subset of links pointing to pages on a specific topic. In the "cooking recipe" case, the performances cannot be improved even using page content, since all the considered sites are effectively on the topic "cooking recipes", and then there is a semantic reason because such sites are connected. We claim that such behavior is intrinsic of the HITS model.

The PageRank algorithm is not topic dependent. Since some pages are referred to by many Web pages independently from their content, such pages result in always being authoritative, regardless the topic of interest. For example, in figures 2 and 4 pages like "www.yahoo.com", "www.google.com", etc, are shown in the top list even if they are not closely related to the specific topic. It is shown in figures 3 and 5 that the "Focused PageRank" WPSS described in section 4.1 can filter many off-topic authoritative pages from the top list. Finally, the "Double Focused PageRank" WPSS is even more effective in filtering all the off-topic authorities, while pushing all the authorities on the relevant topic to the top positions.

5.3 Evaluating the WPPSs

In order to evaluate the proposed WPSSs, we employed a methodology similar to that one presented in [12]. For each WPSS we selected the N pages with highest score, creating a collection of pages to be evaluated by a pool of humans. 10 experts on the specific topics independently labelled each page in the collection as "authoritative for the topic" or "not authoritative for the topic". Such a reliable set of judgments was finally used to measure the percentage of positive (or negative) judgments on the best N pages returned by each ranking function. In particular, N was varied between 1 and 300. The topics selected for these experiments were

Page Rank
www.bravenet.com
www.yahoo.com
www.ebay.com
chef2chef.com
www.internet.com
www.livve.com
www.ringsurf.com
www.gograph.com

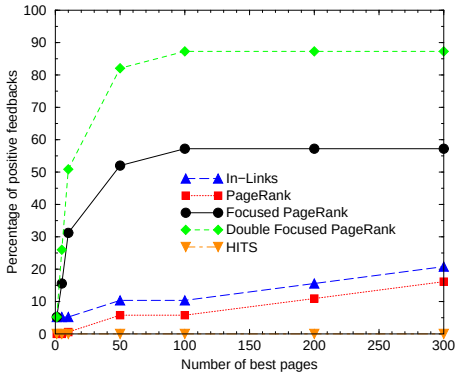
HITS
www.allrecipes.com/default.asp
www.cookierecipe.com
www.vegetarianrecipe.com
www.barbequerecipe.com
seafoodrecipe.com/Default.asp
www.saladrecipe.com
seafoodrecipe.com
www.chickenrecipe.com

Figure 4: We report the 8 top score pages from a portion of the Web, focused on the topic "cooking recipes", using either the PageRank surfer, or the HITS surfer pool. For the HITS surfer pool we report the pages with the top authority value.

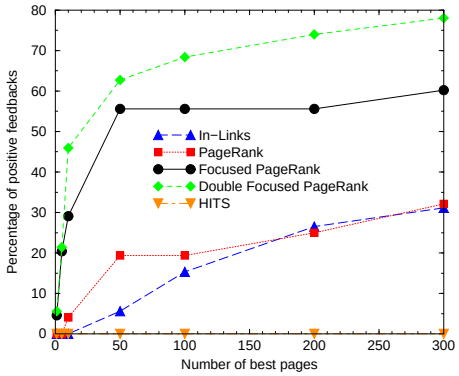
Focused PageRank
www.yahoo.com
chef2chef.com
www.bravenet.com
www.ebay.com
my.lycos.com
www.livve.com
www.netrelief.com
www.freeservers.com

Double Focused PageRank
chef2chef.com
mcgees.com/kitchen
www.bravenet.com
adserve.greekcuisine.com
recipe-cookie.com
www.mega-zine.com/kitchen
www.eaglebrand.com
www.ok-web-design.com

Figure 5: We report the 8 top score pages from a portion of the Web, focused on the topic "cooking recipes", using the proposed focused PageRank surfers.



(a)



(b)

Figure 6: The plots show the percentage of positive judgments assessed by a set of 10 users on the best N pages returned by the various WPSSs respectively for the topic "Linux" (case a) and "Golf" (case b).

"Linux" and "Golf". As in the previous experiments 150.000 pages were collected by focus crawling the Web.

Figure 6 reports the percentage of positive judgments on the N best pages returned by the five WPSSs, respectively, for the topic "Linux" and "Golf". In both cases the HITS algorithm is clearly the worst among the other ones. Since its performance decreases significantly when applied to the entire collection of documents, it can only be used as a query-dependent ranking schema [1].

As previously reported in [12], in spite of its simplicity the In-link algorithm has performances similar to PageRank. In our experiments PageRank outperformed the In-Links algorithm on the category "Golf", whereas it was outperformed on the category "Linux". However, in both cases the gap is small. The two focused ranking functions clearly outperformed all the not focused ones, demonstrating that when searching focused authorities, a higher accuracy is provided by employing a stable computation schema and by taking into account the page content.

6. CONCLUSIONS

In this paper, we have proposed a general framework for the definition of web page scoring systems for horizontal and vertical search engines. The proposed scheme incorporates many relevant scoring models proposed in the literature. Moreover, it contains novel features which looks very appropriate especially for vortals. In particular, the topological structure of the web as well as the content of the web pages play jointly a crucial rule for the construction of the scoring. The experimental results support the effective-

ness of the proposal which clearly emerge especially for vertical search. Finally, it is worth mentioning that the model described in this paper is very well-suited for the construction of learning-based WPSS, which can, in principle, incorporate the user information while surfing the Web.

Acknowledgments
We would like to thank Ottavio Calzone who performed some of the experimental evaluations of the scoring systems.

7. REFERENCES

[1] M. Henzinger, "Hyperlink analysis for the Web," *IEEE Internet Computing*, vol. 1, pp. 45–50, January/February 2001.

[2] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank citation ranking: Bringing order to the web," tech. rep., Computer Science Department, Stanford University, 1998.

[3] J. Kleinberg, "Authoritative sources in a hyperlinked environment." Report RJ 10076, IBM, May 1997, 1997.

[4] K. Bharat and M. Henzinger, "Improved algorithms for topic distillation in a hyperlinked enviroment," in *Proceedings of the 21st ACM SIGIR Conference on Research and Developments in Information Retrieval*, pp. 104–111, 1998.

[5] R. Lempel and S. Moran, "The stochastic approach for link-structure analysis (SALSA) and the TKC effect," in *Proceedings of the 9th World Wide Web Conference*, 2000.

[6] R. Lempel and S. Moran, "Salsa: The stochastic approach for link-structure analysis," *ACM Transactions on Information Systems*, vol. 19, pp. 131–160, April 2001.

[7] D. Cohn and H. Chang, "Learning to probabilistically identify authoritative documents," in *Proc. 17th International Conf. on Machine Learning*, pp. 167–174, Morgan Kaufmann, San Francisco, CA, 2000.

[8] D. Cohn and T. Hofmann, "The missing link: a probabilistic model of document content and hypertext connectivity," in *Neural Information Processing Systems*, vol. 13, 2001.

[9] E. Seneta, *Non-negative matrices and Markov chains*. Springer-Verlag, 1981.

[10] M. Joshi, V. Tawde, and S. Chakrabarti, "Enhanced topic distillation using text, markup tags, and hyperlinks," in *International ACM Conference on Research and Development in Information Retrieval (SIGIR)*, August 2001.

[11] M. Diligenti, F. Coetzee, S. Lawrence, L. Giles, and M. Gori, "Focus crawling by context graphs," in *Proceedings of the International Conference on Very Large DataBases, 11-15 September 2000, Il Cairo, Egypt*, pp. 527–534, 2000.

[12] B. Amento, L. Terveen, and W. Hill, "Does authority mean quality? predicting expert quality ratings of Web documents," in *Proceedings of the 23rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 296–303, 2000.