What is this Page Known for? Computing Web Page Reputations

Davood Rafiei
Department of Computer Science
University of Toronto
drafiei@cs.toronto.edu

Alberto O. Mendelzon
Department of Computer Science
University of Toronto
mendel@cs.toronto.edu

Abstract

The textual content of the Web enriched with the hyperlink structure surrounding it can be a useful source of information for querying and searching. This paper presents a search process where the input is the URL of a page, and the output is a ranked set of topics on which the page has a reputation. For example, if the input is *www.gamelan.com*, then a possible output is "Java." We propose several algorithmic formulations of the notion of reputation using simple random walk models of Web browsing behaviour. We give preliminary test results on the effectiveness of these algorithms.

Keywords: *Reputation Ranking, Searching, Random Walks, PageRank, Hubs and Authorities.*

1 Introduction

The idea of exploiting the "reputation" of a Web page when searching has attracted research attention recently and even been incorporated into some search engines [15, 5, 11, 2, 3]. The idea is that pages with good reputations should be given preferential treatment when reporting the results of a search; and that link structure can be mined to extract such reputation measures, on the assumption that a link from page a to page b is, to some degree, an endorsement of the contents of b by the creator of a.

We consider a different question in this paper: given a page (or a Web site), on what topics is this page considered an authority by the Web community? There are many potential applications for such computations. For example, organizations routinely expend a great deal of effort and money in determining how they are perceived by the public; evaluating the reputation of their Web site on specific topics, or determining those topics on which its reputation is highest (or abnormally low) could be a valuable part of this self-evaluation. A second application is page classification: determining that a page has high reputation on a certain topic is evidence that the page is, first of all, *about* that topic, and also a good candidate to be included in a directory of resources on the topic. Yet another application is the analysis of the reputation of personal home pages to determine what topics a person is known for, say for tenure hearings or recruiting.

However, there are some difficulties in formalizing the concept of "reputation" effectively. The assumption that links are endorsements suggests that the number of incoming links of a page indicates its reputation. But in practice, links represent a wide variety of relationships such as navigation, subsumption, relatedness, refutation, justification, etc. In addition, we are interested not just in the overall reputation of a page, but in its reputation on certain topics. In the next subsection we give an overview of our approach to dealing with these difficulties.

1.1 Overview

We focus on two problems: computing the reputation rank of a page, whether overall or for specific topics; and identifying those topics for which a page has a good reputation. We address these problems in the framework of simple probabilistic models of user behavior that simulate the way pages are created or searched.

We propose two methods for computing the reputations of a page. Our first method is based on one-level weight propagation, generalizing the Page Rank model [3]. The reputation of a page on a topic is proportional to the sum of the reputation weights of pages pointing to it on the same topic. In other words, links emanating from pages with high reputations are weighted more. For example, a page can acquire a high reputation on a topic because the page is pointed to by many pages on that topic, or because the page is pointed to by some high reputation pages on that topic.

Our second method is based on two-level weight propagation, generalizing the Hubs and Authorities model [11]. In this model, a page is deemed an *authority* on a topic if it is pointed to by good *hubs* on the topic, and a good hub is one that points to good authorities.

We formulate both these methods in terms of random walks on the Web graph. Our random walk formulation of the first method is an extension of the one used to define PageRank [3]; unlike PageRank our formulation allows computing the reputation rank of a page on a specific topic. Our random walk formulation of the second method is novel; to the best of our knowledge, there is no random walk formulation of a hubs-and-authorities model in the literature. We present algorithms for computing page reputations both in the case where a large crawl of the Web is available and when it is not. We also provide preliminary experimental results on the effectiveness of our formulations.

1.2 Related Work

Recent work on analyzing the link structure of the Web suggests that hyperlinks between pages often represent relevance [15, 5] or endorse some authority [11, 2, 3].

Brin and Page [3] suggest a recursive method for ranking the importance of a Web page based on the importance of its incoming links. The ranking is based on simulating the behavior of a "random surfer" who either selects an outgoing link uniformly at random, or jumps to a new page chosen uniformly at random from the entire collection of pages. The PageRank of a page corresponds to the number of visits the "random surfer" makes to the page. The Google search engine [9] adopts PageRank as part of its ranking system. Our first model of ranking is an extension of PageRank; the main difference is we do ranking with respect to a topic instead of computing a universal rank for each page.

Kleinberg [11] proposes an algorithm that, given a topic, finds pages that are considered strong authorities on that topic. For example, given the term "Java", the system built around this algorithm, known as HITS, finds *www.gamelan.com* among other pages. The algorithm is based on the intuition that for broad topics, authority is conferred by a set of *hub pages*, which are recursively defined as a set of pages with a large number of links to many relevant authorities. The basic idea is to compile a root set of pages that contain the query terms, extend this set by adding pages linked to/from these pages, build the adjacency matrix A of the link graph, and compute the eigenvectors of A^TA and AA^T. These vectors respectively correspond to the weights of authorities and hubs. We provide a probabilistic formulation of this search mechanism which also allows us to go in the opposite direction, i.e. given the URL of a page, we can find the topics the page is an authority on.

The literature reports analyses and improvements over Kleinberg's original algorithm. Gibson et al. [8] investigate the dependence between top authorities and hubs identified by HITS and the choice of the root set. Bharat and Henzinger [2] suggest the use of link weights to adjust the influence of pages based on their relevance to the query. To measure the relevance of a page to a query, they use the normalized *cosine* measure of similarity between the page and an estimated query page, computed by concatenating the first 1000 words of pages retrieved in the root set.

Based on the hub-and-authority structure of a community, Kumar et al. [13] show that a large number of such communities can be identified from their signatures in the form of complete bipartite subgraphs of the Web. Chakrabarti, Dom, and Indyk [4] show the benefit of using linkage information within a small neighborhood of documents to improve the accuracy of a text-based statistical classifier. Dean and Henzinger [5] suggest algorithms to find related pages of a given page solely based on the linkage structure around the page. Finally, Henzinger et al. [10] use random walks on the Web to measure the quality of pages stored in an index.

The view of the Web as a directed-graph database allows a large number of database techniques to be applied to the Web. Several query languages have been proposed for both querying and restructuring Web documents. A recent survey by Florescu, Levy and Mendelzon [7] gives an overview of this area.

2 Random Walks on the Web

Given a set $S = \{s_1, s_2, \ldots, s_n\}$ of states, a *random walk* on S corresponds to a sequence of states, one for each step of the walk. At each step, the walk either 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 it only depends on the current state. A random walk on the Web is in the form of navigation between pages, where each page represents a possible state, and each link represents a possible transition.

2.1 One-Level Influence Propagation

Consider a "random surfer" who wanders the Web, searching for pages on topic t. At each step, the surfer either jumps into a page uniformly chosen at random from the set of pages that contain the term t, or follows a link uniformly chosen at random from the set of outgoing links of the current page. If the random surfer continues this walk forever, then the number of visits he or she makes to a page is its reputation on t.

Intuitively, pages with relatively high reputations on a topic are more likely to be visited by the random surfer searching for that topic. A justification for this is that the reputation of a page on a topic naturally depends both on the number of pages on the same topic that point to it, and on the reputations of these pages on the same topic as well. The number of visits the surfer makes to a page depends on the same two factors.

2.1.1 Formal Model

We want to *define* the reputation of a page p on topic t as the probability that the random surfer looking for topic t will visit page p. For this we formulate the following random walk model.

Suppose at each step, with probability d the random surfer jumps into a page uniformly chosen at random from the set of pages that contain the term t, and with probability $(1-d)$ he or she follows an outgoing link from the current page. Let N_t denote the total number of pages on the Web that contain the term t. Intuitively, the probability that the surfer at each step visits page p in a random jump is d/N_t if page p contains term t and it is zero otherwise. Let $q \to p$ denote a link from page q to page p, and $O(q)$ denote the number of outgoing links of page q. Intuitively, the probability that the surfer visits page p at step n after visiting page q and through the link $q \to p$ is $((1-d)/O(q))R^{n-1}(q,t)$ where $R^{n-1}(q,t)$ denotes the probability that the surfer visits page q for topic t at step $n-1$. We can write the probability of visiting page p for topic t at step n of the walk as follows:

$$R^n(p,t) = \begin{cases} \dfrac{d}{N_t} + (1-d) \sum\limits_{q \to p} \dfrac{R^{n-1}(q,t)}{O(q)} & \textit{if term } t \textit{ appears in page } p \\ (1-d) \sum\limits_{q \to p} \dfrac{R^{n-1}(q,t)}{O(q)} & \textit{otherwise} \end{cases} \tag{1}$$

Definition 1 *The one-level reputation rank of page p on topic t is the equilibrium probability $\pi_{p,t}$ of visiting page p for topic t, i.e.*

$$\pi_{p,t} = \lim_{n \to \infty} R^n(p,t). \tag{2}$$

Theorem 1 *The notion of one-level reputation rank is well-defined, i.e. for every term t with $N_t > 0$ and every parameter $d > 0$, there is a unique probability distribution $\pi_{p,t}$ satisfying Equation 2, provided that every page has at least one outgoing link.*

Proof: Given a term t and a parameter $d > 0$, consider the base set of pages that contain the term t, and add to this set every page which can be reached from a page in the base set. Construct the matrix U of transition probabilities for the random walk process with each entry u_{ij} representing the probability of directly going from page p_i to page p_j as follows: first, if there is no link from p_i to p_j, then set entry u_{ij} of the matrix to 0, otherwise set it to $(1-d)/O(p_j)$; second, add d/N_t to every entry u_{ij} where p_j contains the term t. Clearly U is a square stochastic matrix with non-negative elements and unit row sums due to the assumption that every page has at least one outgoing link. Thus, both U and U^T have eigenvectors with eigenvalue 1. If we denote the weights of pages in the current step of the walk with vector $\vec{x}$, then the weights in the next step of the walk will be $\vec{x} = U^T\vec{x}$. Therefore, we are seeking an eigenvector of U associated with the eigenvalue 1.

Furthermore, because of the parameter $d > 0$, the transition matrix is both *irreducible* (i.e., every page is reachable from every other page) and *aperiodic* (see, for example [16], for details). Therefore, according to the convergence theorem [16, Theorem 1.8.3], starting from any distribution $\vec{x}$, $(U^T)^{(n)}\vec{x}$ will converge to the stationary probability $\pi_{p,t}$ of pages induced by the random walk process when $n \to \infty$. ∎

In the setting of the Web, our assumption that every page has at least one outgoing link may not be true; there are often pages that have no outgoing link, or the outgoing links may not be valid. A solution to accommodate these pages is to implicitly add links from every such page to all pages in the base set, i.e. the set of pages that contain the term. The interpretation here is that when the surfer reaches a dead end, he or she jumps to a page in the base set chosen uniformly at random.

2.2 Two-Level Influence Propagation

We return to the "random surfer" who wanders the Web, searching for pages on topic t. The surfer's behaviour is a bit more involved now. Define a *transition* as one of (a) jump to a page on topic t chosen uniformly at random from the whole collection, or (b) follow an outgoing link of the current page chosen uniformly at random. When the current page is p, the surfer has two choices: either make a transition out of page p, or randomly pick any page q that has a link into page p and make a transition out of page q. The intuitive justification is this: when the surfer reaches a page p that seems useful for topic t, this does not mean that p is a good source of further links; but it does mean that pages q that point to p may be good sources of links, since they already led to page p.

To make our presentation slightly more formal, we say the surfer follows links both *forward* (out of page p) and *backward* (into page q). The walk alternates strictly between forward and backward steps, except that after option (a) is chosen, the direction of the next step is picked at random.

If the random surfer continues the walk forever, then the number of forward visits he or she makes to a page is its *authority* reputation and the number of backward visits he or she makes to a page is its *hub* reputation. Clearly pages with relatively high authority reputations on a topic are more likely to be visited through their incoming links, and pages with relatively high hub reputations on a topic are more likely to be visited through their outgoing links. Intuitively the authority reputation of a page p on topic t depends not only on the number of pages on topic t that point to p, but on the hub reputations of these pages on topic t as well. Similarly, the hub reputation of a page p on topic t depends not only on the number of pages on topic t that page p points to, but on the authority reputations of these pages on topic t as well.

2.2.1 Formal Model

We want to *define* the *authority* reputation of a page p on a topic t as the probability that the random surfer looking for topic t makes a forward visit to page p and the *hub* reputation of a page p on topic t as the probability that the random surfer looking for topic t makes a backward visit to page p. For this we formulate the following random walk model.

Suppose at each step, with probability d the random surfer picks a direction and jumps into a page uniformly chosen at random from the set of pages on topic t, and with probability $(1-d)$ the surfer follows a link. Intuitively, the probability that at each step the surfer makes a forward visit (and similarly a backward visit) to page p in a random jump is $d/2N_t$ if page p contains term t and it is zero otherwise. Let $p \rightarrow q$ denote a link from page p to page q, $O(p)$ denote the number of outgoing links of page p, and $I(p)$ denote the number of incoming links of page p. Let us denote with $A^{n-1}(p,t)$ the probability of a forward visit into page p at step $n-1$ and with $H^{n-1}(p,t)$ the probability of a backward visit into page p at step $n-1$. Intuitively, the probability that the surfer makes a forward visit to page p at step n after visiting page q and through a link $q \rightarrow p$ is $((1-d)/O(q))H^{n-1}(q,t)$. Similarly, the probability that the surfer makes a backward visit to page q at step n after visiting page p and through a link $q \rightarrow p$ is $((1-d)/I(p))A^{n-1}(p,t)$. We can write the probabilities, $A^n(p,t)$ and $H^n(p,t)$, of visiting page p for topic t at step n as follows:

$$A^n(p,t) = \begin{cases} \dfrac{d}{2N_t} + (1-d)\displaystyle\sum_{q \rightarrow p} \dfrac{H^{n-1}(q,t)}{O(q)} & \textit{if term } t \textit{ appears in page } p \\ (1-d)\displaystyle\sum_{q \rightarrow p} \dfrac{H^{n-1}(q,t)}{O(q)} & \textit{otherwise} \end{cases} \tag{3}$$

$$H^n(p,t) = \begin{cases} \dfrac{d}{2N_t} + (1-d)\displaystyle\sum_{p \rightarrow q} \dfrac{A^{n-1}(q,t)}{I(q)} & \textit{if term } t \textit{ appears in page } p \\ (1-d)\displaystyle\sum_{p \rightarrow q} \dfrac{A^{n-1}(q,t)}{I(q)} & \textit{otherwise} \end{cases} \tag{4}$$

Definition 2 *The two-level reputation rank $r \in \{authority, hub\}$ of page p on topic t is the equilibrium probability $\pi^r_{p,t}$ of visiting page p for topic t in the direction associated to r (forward for authority and backward for hub), i.e.*

$$\pi^{authority}_{p,t} = \lim_{n \rightarrow \infty} A^n(p,t) \tag{5}$$

$$\pi^{hub}_{p,t} = \lim_{n \rightarrow \infty} H^n(p,t). \tag{6}$$

Theorem 2 *The notion of two-level reputation rank is well-defined, i.e. for every term t with $N_t > 0$ and every parameter $d > 0$, there is a unique probability distribution $\pi^r_{p,t}$ satisfying Equations 5 and 6, if every page has at least an incoming and an outgoing link.*

Proof: Given a term t and a parameter $d > 0$, consider the base set of pages that contain the term t, and add to this set every page which is reachable from a page in the base set by repeatedly following links in one of the *back-forth* or the *forth-back* order. To construct the matrix U of transition probabilities for the random walk process, we allocate two states for each page p_i, say i^f to denote the state of the page when it is visited in the forward direction and i^b to denote the state of the page when it is visited in the backward direction. Entries of matrix U are set as follows: (1) $u_{i^f j^f} = u_{i^b j^b} = 0$; (2) if there is a link from page p_i to page p_j, then $u_{i^b j^f} = (1-d)/O(p_i)$ and $u_{j^f i^b} = (1-d)/I(p_j)$; otherwise $u_{i^b j^f} = u_{j^f i^b} = 0$; (3) add $d/2N_t$ to every entry in columns j^f and j^b if p_j is on topic t.

Clearly U is a square stochastic matrix with non-negative elements and unit row sums due to the assumption that every page has at least an incoming and an outgoing link. Thus, both U and U^T have eigenvectors with eigenvalue 1. If we denote the weights of pages in the current step of the walk with vector $\vec{x}$, then the weights in the next step of the walk will be $\vec{x} = U^T \vec{x}$. Therefore, we are seeking an eigenvector of U associated with the eigenvalue 1.

The transition matrix U is both irreducible and aperiodic; therefore, according to the convergence theorem [16, Theorem 1.8.3], starting from any distribution $\vec{x}$, $(U^T)^{(n)}\vec{x}$ will converge to the stationary probability $\pi^r_{p,t}$ of pages induced by the random walk process when $n \to \infty$. ∎

In the setting of the Web, our assumption that a page has at least one incoming link and one outgoing link may not hold. One solution to accomodate such pages is to assign a hub (or authority) rank of zero for every page with no outgoing (incoming) links. The random walk process accordingly can be modified by restricting random jumps only to pages with at least either one incoming link or one outgoing link. Another solution is to add a link from every page to itself. This ensures that every page will acquire a fixed minimum authority and hub rank on topics of the page independent from its links.

3 Computing Reputations of Pages

The probabilistic models presented in the previous section provide a natural way of measuring the reputations of a page, but there are computational issues which need to be addressed. The first issue is within which set of pages should the ranks be computed. The second issue is what is the set of topics on which to compute reputations. It is not enough to look for terms or phrases that appear in a page, as a page might have a high reputation on a topic, but the term denoting that topic may not be explicitly mentioned anywhere in the page. For example, Sun Microsystems has a high reputation on "Java", but the term does not appear in *www.sun.com*. In this section, we address both problems. Subsection 3.1 deals with the situation where we have access to a large crawl of the Web, as is the case for example when the computation is performed by a search engine. Subsection 3.2 deals with the situation where we do not have access to such a crawl or cannot afford the time to do the full computation of subsection 3.1.

3.1 Computing Reputation Ranks

Given a collection of pages, for example the result of a relatively large crawl of the Web, and a parameter d, we can compute the reputation ranks using one of the two influence propagation models. The ranks in the one-level influence propagation model are in the form of a sparse matrix, say R, with rows representing Web pages and columns denoting each term or phrase that appears in some document (after removing stop words, etc.) The computation involves initializing R and repeatedly updating it until convergence.

Algorithm 1 : Computing One-Level Reputation Ranks

For every page p and term t,
  Initialize $R(p,t) = 1/N_t$ if t appears in page p; otherwise $R(p,t) = 0$.
While R has not converged,
  Set $R'(p,t) = 0$ for every page p and term t,
  For every link $q \rightarrow p$,
    $R'(p,t) = R'(p,t) + R(q,t)/O(q)$
  $R(p,t) = (1-d)R'(p,t)$ for every page p and term t,
  $R(p,t) = R(p,t) + d/N_t$ if term t appears in page p.

Since each column of R converges to the principal eigenvector of the matrix of transition probabilities for a term t, the algorithm is guaranteed to converge. The principal eigenvector associated to each term is the stationary distribution of pages in the random walk process, provided every page has at least one outgoing link and $d > 0$.

The ranks in the two-level influence propagation model can be represented in the form of two sparse matrixes, say H and A, respectively denoting the hub and the authority reputations of pages. The computation can be arranged as follows:

Algorithm 2 : Computing Two-Level Reputation Ranks

For every page p and term t,
  Initialize $H(p,t) = A(p,t) = 1/2N_t$ if t appears in page p; otherwise $H(p,t) = A(p,t) = 0$.
While both H and A have not converged,
  Set $H'(p,t) = A'(p,t) = 0$ for every page p and term t,
  For every link $q \rightarrow p$,
    $H'(q,t) = H'(q,t) + A(p,t)/I(p)$
    $A'(p,t) = A'(p,t) + H(q,t)/O(q)$
  $H(p,t) = (1-d)H'(p,t)$ and $A(p,t) = (1-d)A'(p,t)$ for every page p and term t,
  $H(p,t) = H(p,t) + d/2N_t$ and $A(p,t) = A(p,t) + d/2N_t$ if term t appears in page p.

Again, the computation for each term is guaranteed to converge to the principal eigenvector of the matrix of transition probabilities for that term. The principal eigenvector is the stationary distribution provided every page has at least one incoming or outgoing link and $d > 0$. Next we discuss how to obtain an approximate estimation of reputation when we do not have access to a large crawl of the Web.

3.2 Identifying Topics

The two algorithms presented in the previous section not only compute the reputation ranks but also identify topics of reputations. However, in practice we may not have access to a large crawl of the Web, or we may not be able to afford the full computation. In this section, we show that it is still possible to approximately find the topics a page has a high reputation on, although the ranks will not reflect the real probability distributions.

Given a page p and a parameter $d > 0$, suppose we want to find the reputations of the page within the one-level influence propagation model. If the page acquires a high rank on an arbitrarily chosen term t within the full computation of Algorithm 1, then at least one of the following must hold: (1) term t appears in page p, (2) many pages on topic t point to p, or (3) there are pages with high reputations on t that point to p. This observation provides us with a practical way of identifying the candidate terms. We simply start from page p and collect all terms that appear in it. We then look at the incoming links of the page and collect all possible terms from those pages. We continue this process until we get to a point where either there is no incoming link or the incoming links have very small effects on the reputations of page p. Let us denote the maximum number of iterations by k. The algorithm can be expressed as follows:

Algorithm 3 : Approximating One-Level Reputation

$R(p,t) = d/N_t$ for every term t that appears in p
For l = 1,2, …, k
  $d' = d$ if $l < k$, 1 otherwise
  For every path $q_l \rightarrow \ldots \rightarrow q_1 \rightarrow p$ of length l and every term t in page q_l,
    $R(p,t) = 0$ if term t has not been seen before
    $R(p,t) = R(p,t) + ((1-d)^l / \prod_{i=1}^{l} O(q_i))(d'/N_t)$
Report every term t with $R(p,t) > 1/N_t$.

The parameter k can be chosen such that $(1-d)^k$ becomes very close to zero; i.e. there is no need to look at a page if the terms that appear in the page have little or no effect in the reputations of page p. Similarly, the hub and the authority reputations of a page can be approximated within the two-level influence propagation model as follows:

Algorithm 4 : Approximating Two-Level Reputation

$H(p,t) = A(p,t) = d/(2N_t)$ for every term t that appears in p
For l = 1,2, …, k
  $d' = d$ if $l < k$, 1 otherwise
  If l is odd
    For every path $q_l \rightarrow q_{l-1} \leftarrow q_{l-2} \ldots \rightarrow p$ of length l and every term t in page q_l,
      $A(p,t) = 0$ if term t has not been seen before
      $A(p,t) = A(p,t) + ((1-d)^l/(O(q_l)I(q_{l-1})\ldots O(q_1)))d'/(2N_t)$
    For every path $p \rightarrow \ldots q_{l-2} \leftarrow q_{l-1} \rightarrow q_l$ of length l and every term t in page q_l,
      $H(p,t) = 0$ if term t has not been seen before
      $H(p,t) = H(p,t) + ((1-d)^l/(I(q_l)O(q_{l-1})\ldots I(q_1)))d'/(2N_t)$
  else
    For every path $q_l \leftarrow q_{l-1} \rightarrow \ldots \rightarrow p$ of length l and every term t in page q_l,
      $A(p,t) = 0$ if term t has not been seen before
      $A(p,t) = A(p,t) + ((1-d)^l/(I(q_l)O(q_{l-1})\ldots))d'/(2N_t)$
    For every path $p \rightarrow \ldots \rightarrow q_{l-1} \leftarrow q_l$ of length l and every term t in page q_l,
      $H(p,t) = 0$ if term t has not been seen before
      $H(p,t) = H(p,t) + ((1-d)^l/(O(q_l)I(q_{l-1})\ldots))d'/(2N_t)$

Report every term t with $A(p,t) > 1/2N_t$ or $H(p,t) > 1/2N_t$.

In both algorithms 3 and 4, we have adopted a breadth-first search of the pages that can affect the reputations of a page p, i.e. all pages within depth l are visited before any page in depth $l+1$. A benefit of this ordering is that the user can stop the search at any point and be sure that pages that are expected to have a high influence on p are visited. This may happen, for example, if the search takes longer than expected. However, it should be noted that the algorithm needs to remember the number of outgoing or incoming links for each page being visited, if this information is not already stored. An alternative to a breadth-first search is to conduct a depth-first search, if we can assume that, for example, the search engine always gives us the same set of pages with the same ordering. The only benefit of such a search is that the algorithm only needs to remember the current path. However, this assumption usually does not hold for real search engines. In addition, there is the danger of spending most of the time on pages that have a very small effect on the reputations of page p before visiting more important pages.

4 Duality of Terms and Pages

Our main objective so far has been to find the topics on which a page has a strong reputation, but our random walk models also allow us to compute the pages that have high reputation on a given topic, as proposed by Kleinberg and others for enhancing search engine performance.

Indeed, if we fix p in equations 1, 3, 4 to a specific page, we will find the reputation ranks of the page for every possible topic t. We may then report the topics with the highest reputation ranks. If we fix instead t in the same equations to be a specific topic, we will find the reputation ranks of every page on topic t. Again, we may report those pages with high reputation ranks first in the answer to a query.

In terms of rank computations, our algorithms presented in Section 3.1 already compute the reputations of every page p on every topic t. Therefore, the highly-weighted pages for a given topic can be easily identified. In practice, however, we may not be able to afford the full computation for every possible term; or an approximate solution might be as good as an exact solution. In Section 3.2 we presented algorithms to approximately find the topics on which a page has a high reputation. In the rest of this section, we show how we can approximately find pages with relatively high reputations on a given topic.

Given a topic t, an arbitrarily chosen page p can potentially acquire a relatively high rank, within the one-level influence propagation model, on topic t if at least one of the following hold: (1) term t appears in page p, (2) many pages on topic t point to p, (3) there are pages with relatively high reputations on t that point to p. Thus, a page with high reputation on topic t must either contain term t or be reachable within a few steps from a large set of pages on topic t. An approximate way of computing the one-level reputation ranks of pages on topic t is as follows: (1) identify pages that are either on topic t or reachable within a short distance from a page on topic t; (2) construct the matrix U of transition probabilities for the resulting set of pages, as described in Section 2.1; (3) compute the principal eigenvector of U^T. The principal eigenvector will give the approximate ranks of pages that are expected to have high reputations; i.e. every page which is not identified in Step 1 is assumed to have a rank of zero. This is more general than the PageRank computation, which determines the *overall* reputation of a page, but not its reputation on specific topics.

For the two-level influence propagation model, given a topic t, an arbitrarily chosen page p can acquire a relatively high rank on topic t if either term t appears in page p or it is reachable within a short path of alternating forward and backward links (or vice versa) from a large set of pages on topic t. An approximate way of computing the two-level reputation ranks of pages on topic t is as follows: (1) identify pages that are either on topic t or reachable within few steps from a page on topic t, alternately following links forward and backward or vice versa; (2) construct the matrix U of transition probabilities for the resulting set of pages, as described in Section 2.2; (3) compute the principal eigenvector of U^T. The principal eigenvector will give the approximate ranks of pages that are expected to have high reputations. Again, every page which is not identified in Step 1 is assumed to have a rank of zero. Note that the hubs-and-authorities computation of Kleinberg is a special case of this method; it is based on only identifying pages that either contain term t or are reachable within one link from one such page.

5 Experimental Evaluation

In this section, we describe a preliminary evaluation of our approach. Since we did not have access to a large crawl of the Web, it was not feasible to do the full rank computations of Section 3.1. We also did not fully implement the approximate algorithms suggested in Section 3.2 due to the limitations imposed by the search engines we used, either on the maximum number of entries returned for a query or on the response time.

Instead, we implemented a simplified version of Algorithm 3 (and also part of Algorithm 4 that computes the authority reputation of a page) where we set k to 1, d to 0.10 and $O(q_i)$ for every page q_i to 7.2, the estimated average number of outgoing links of a page [12]. The best value for parameter d needs to be determined empirically. Further details of the implementation are as follows:

1. Only a limited number of incoming links are examined; we obtain at most 500 incoming links of a page, but the number of links returned by the search engine, currently Alta Vista [1], can be less than that.

2. For each incoming link, terms are extracted from the "snippet" returned by the search engine, rather than the page itself. A justification for this is that the snippet of a page, to some degree, represents the topic of the page. In addition, the number of distinct terms and as a result the number of count queries needed to be sent to the search engine are dramatically reduced.

3. Internal links and duplicate snippets are removed.

4. Stop words and every term t with $N_t < (1+r*L)$ are removed, where L is the number of incoming links collected and r is the near-duplicate ratio of the search engine, currently

URL : java.sun.com   500 links examined (out of 128653 available) **Highly weighted terms**: Developers, JavaSoft, Applets, JDK, Java applets, Sun Microsystems, API, Programming, Solaris, tutorial **Frequent terms**: Java, Software, Computer, Programming, Sun, Development, Microsoft, Search
URL : sunsite.unc.edu/javafaq/javafaq.html   500 links examined (out of 1541 available) **Highly weighted terms**: Java FAQ Java, comp.lang.java FAQ, Java Tutorials, Java Stuff, Applets, IBM Java, Javasoft, Java Resources, API Java, Learning Java **Frequent terms**: Java, Programming, FAQ, Sun, Computer, Language, Tutorial, Java FAQ, Software

Figure 1: Authorities on (java)

set to 0.01. This reduces the number of count queries and also removes unusual terms such as "AAATT" that rarely appear in any page but might acquire high weights.

Despite all the simplifications, the experience with our prototype has been quite encouraging in terms of approximating both the one-level reputation and the two-level authority reputation of a page. Next we report our experiments with the prototype, called *TOPIC*, which can be tried online at `http://www.cs.toronto.edu/db/topic`.

5.1 Known Authoritative Pages

In our first experiment, we picked a set of known authoritative pages on queries *(java)* and *(+censorship +net)*, as reported by Kleinberg's HITS algorithm [11], and computed the topics that each page was an authority on. As shown in Figure 1, the term "java" is the most frequent term among pages that point to an authority on Java. There are other frequent terms such as "search" or "Microsoft" which have nothing to do with the topic; their high frequency represents the fact that authorities on Java are frequently cocited with search engines or Microsoft. This usually happens in cases where the number of links examined is much less than the number of links available. However, the highly-weighted terms for each page in both Figures 1 and 2 largely describe the topics that the page is an authority on consistently with the results of HITS.

In another experiment, we used Inquirus [14], the NECI meta-search engine, which computes authorities using an unspecified algorithm. We provided Inquirus with the query *("data warehousing")* and set the number of hits to its maximum, which was 1,000, to get the best authorities, as suggested by the system. We picked the top four authorities returned by Inquirus and used our system to compute the topics those pages have high reputations on. The result, as shown in Figure 3, is again consistent with the judgments of Inquirus.

URL : www.eff.org   500 links examined (out of 181899 available) **Highly weighted terms**: Anti-Censorship, Join the Blue Ribbon, Blue Ribbon Campaign, Electronic Frontier Foundation, Free Speech
URL : www.cdt.org   500 links examined (out of 12922 available) **Highly weighted terms**: Center for Democracy and Technology, Communications Decency Act, Censorship, Free Speech, Blue Ribbon, Syllabus, encryption
URL : www.vtw.org   500 links examined (out of 7948 available) **Highly weighted terms**: decision is near in the fight to overturn the Communications Decency Act, Blue Ribbon Campaign, Censorship, American Civil Liberties Union, free speech
URL : www.aclu.org   500 links examined (out of 22087 available) **Highly weighted terms**: ACLU, American Civil Liberties Union, Communications Decency Act, Amendment, CDA, Criminal Law, Censorship

Figure 2: Authorities on (+censorship +net)

URL : www.dw-institute.com   390 links examined (out of 785 available) **Highly weighted terms**: TDWI, Data Warehousing Information Center, www.dw-institute.com, Data Warehousing Institute, data warehouse
URL : pwp.starnetinc.com/larryg   500 links examined (out of 1017 available) **Highly weighted terms**: Data Warehousing Information Center, OLAP and Data, Analytical Processing, Data Mining, data warehouse, Decision Support Systems
URL : www.datawarehousing.com   188 links examined (out of 229 available) **Highly weighted terms**: Data Warehousing Information, OLAP, Data Mining
URL : www.dmreview.com   270 links examined (out of 1258 available) **Highly weighted terms**: Data Warehouse 100, Powell Publishing, Review Magazine, Data Warehousing, Business Intelligence, Cognos, Data Mining, Product Review

Figure 3: Authorities on ("data warehousing")

URL : www.w3.org/People/Berners-Lee   500 links examined (out of 933 available) **Highly weighted terms**: History Of The Internet, Tim Berners-Lee, Internet History, W3C
URL : www-cs-faculty.stanford.edu/˜ knuth   500 links examined (out of 1733 available) **Highly weighted terms**: Don Knuth, Donald E Knuth, TeX, Dilbert Zone, Latex, ACM
URL : www-db.stanford.edu/˜ ullman   238 links examined (out of 466 available) **Highly weighted terms**: Jeffrey D Ullman, Database Systems, Database Management, Data Mining, Programming Languages, Computer Science, Stanford University
URL : www.cs.toronto.edu/˜ mendel   139 links examined (out of 259 available) **Highly weighted terms**: Alberto Mendelzon, Data Warehousing and OLAP, SIGMOD, DBMS

Figure 4: Personal home pages

5.2 Personal Home Pages

In another experiment, we selected a set of personal home pages and used our system to find the high reputation topics for each page. We expected this to describe in some way the reputation of the owner of the page. The results, as shown in Figure 4, can be revealing, but need to be interpreted with some care. Tim Berners-Lee's reputation on the "History of the Internet," Don Knuth's fame on "TeX" and "Latex" and Jeff Ullman's reputation on "database systems" and "programming languages" are to be expected. The humour site *Dilbert Zone* [6] seems to be frequently cited by Don Knuth's fans. Alberto Mendelzon's high reputation on "data warehousing," on the other hand, is mainly due to an online research bibliography he maintains on data warehousing and OLAP in his home page, and not to any merits of his own.

5.3 Unregulated Web Sites

In our last experiment, we selected the home pages of a number of Computer Science Departments on the Web. The main characteristic of these pages is that the sites are unregulated, in the sense that users store any documents they desire in their own pages. The results, as shown in Figure 5, can be surprising. The Computer Science Department at the University of Toronto has a high reputation on "Russia" and "Images," mainly because a Russian graduate student of the department has put online a large collection of images of Russia, and many pages on Russia link to it. The high reputation on "hockey" is due to a former student who used to play on the Canadian national women's hockey team. The Faculty of Mathematics, Computer Science, Physics and Astronomy at the University of Amsterdam (`www.wins.uva.nl`) has a high reputation on "Solaris 2 FAQ" because the site maintains a FAQ on the Solaris operating system. It also has a high reputation on the musician Frank Zappa because it has a set of pages dedicated to him and the FAQ of the `alt.fan.frank-zappa` newsgroup. The Computer Science Department of the University of Helsinki (`www.cs.helsinki.fi`) has a high reputation on Linux because of the many pages

URL : www.cs.toronto.edu *500 links examined (out of 7814 available)* **Highly weighted terms**: Russia, Computer Vision, Linux, Images, Orthodox, Hockey	
URL : www.wins.uva.nl *500 links examined (out of 6174 available)* **Highly weighted terms**: Solaris 2 FAQ, Wiskunde, Frank Zappa, FreeBSD, Recipes	
URL : www.cs.helsinki.fi *500 links examined (out of 9664 available)* **Highly weighted terms**: Linux Applications, Linux Gazette, Linux Software, Knowledge Discovery, Linus Torvalds, Data Mining	

Figure 5: Computer Science Departments

on Linux that point to Linus Torvalds's page.

5.4 Limitations

There are a number of factors that affect our page reputation computations. The first factor is how well a topic is represented on the Web. A company, for instance, may have a high reputation on a specific topic, or a person may be well known for his or her contribution in a specific field, but their home pages may not receive the same recognition mainly because the topic or the field is not well represented on the Web; or even if it is, it may not be visible among other dominant topics. This can be easily seen in some of our experiments.

The second factor is how well pages on a topic are connected to each other. There are two extreme cases that can affect the convergence of a topic in our computations. At one extreme, there are a few pages such as the Microsoft home page (`www.microsoft.com`) with incoming links from a large fraction of all pages on the Web. These pages end up having high reputation on almost every topic represented in the Web; it is not reasonable to identify a small set of highly-weighted topics for them.

At the other extreme, there are pages with no more than a few incoming links; according to some estimates (e.g. [13]), a large number of pages fall in this category. Depending on where the incoming links of a page are coming from and the reputations of those links, they can have various effects on the reputation of a page according to our models. Our current implementation, however, may not report any strong reputations on any topic for these pages because all incoming links are simply weighted equally.

6 Conclusions

We have introduced general notions of page reputation on a topic, combining the textual content and the link structure of the Web. Our notions of reputation are based on random walk models that generalize the pure link-based ranking methods developed earlier. For instance, our ranking based on the one-level weight propagation model becomes PageRank if the rank is computed with respect to all possible topics. We have presented algorithms for identifying the topics that a page has highest reputation on and for computing the reputation rank of a page on a topic. Our current work concentrates on refining the implementation of TOPIC to achieve more accurate rankings and better performance.

Acknowledgments

This research was supported by Communications and Information Technology Ontario and the Natural Sciences and Engineering Research Council.

References

[1] Alta Vista. http://www.altavista.com.

[2] K. Bharat and M. R. Henzinger. Improved algorithms for topic distillation in hyperlinked environments. In *Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 104–111, 1998.

[3] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the 7th International World Wide Web Conference*, pages 107–117, Brisbane, Australia, April 1998. Elsevier Science.

[4] S. Chakrabarti, B. Dom, and P. Indyk. Enhanced hypertext categorization using hyperlinks. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 307–318, Seattle,WA, 1998.

[5] J. Dean and M. R. Henzinger. Finding related pages on the Web. In *Proceedings of the 8th International World Wide Web Conference*, pages 389–401, Toronto, Canada, May 1999. Elsevier Science.

[6] Dilbert Zone. http://www.unitedmedia.com/comics/dilbert.

[7] D. Florescu, A. Levy, and A. Mendelzon. Database techniques for the World Wide Web : a survey. *ACM SIGMOD Record*, 27(3):59–74, September 1998.

[8] D. Gibson, J. M. Kleinberg, and P. Raghavan. Inferring Web communities from link topology. In *Hypertext*, pages 225–234, Pittsburgh,PA, June 1998.

[9] Google. http://www.google.com.

[10] M. R. Henzinger, A. Heydon, M. Mitzenmacher, and M. Najork. Measuring index quality using random walks on the Web. In *Proceedings of the 8th International World Wide Web Conference*, pages 213–225, Toronto, Canada, May 1999. Elsevier Science.

[11] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. In *Proceedings of ACM-SIAM Symposium on Discrete Algorithms*, pages 668–677, January 1998.

[12] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Extracting large-scale knowledge bases from the Web. In *Proceedings of the 25th International Conference on Very Large Databases*, pages 639–650, September 1999.

[13] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Trawling the Web for emerging cyber-communities. In *Proceedings of the 8th International World Wide Web Conference*, pages 403–415, Toronto, May 1999. Elsevier Science.

[14] S. Lawrence and C.L. Giles. Context and page analysis for improved Web search. *IEEE Internet Computing*, 2(4):38–46, 1998.

[15] Netscape Communications Corporation. What's related. Web page, http://www.netscape.com/escapes/related/faq.html.

[16] J. R. Norris. *Markov Chains*. Cambridge University Press, 1997.