Computer Networks
Elsevier Science B.V., P.O. Box 1930,
1000 BX Amsterdam

Pages: 1 to 15

IN ALL CORRESPONDENCE
CONCERNING THIS PAPER
REFER TO:

COMPNW 2288 / WWW 175

AUTOPAGE PROOF

Computer Networks 00 (2000) COMPNW2288

The stochastic approach for link-structure analysis (SALSA) and the TKC effect [1]

R. Lempel *, S. Moran

Department of Computer Science, The Technion, Haifa 32000, Israel

Abstract

Today, when searching for information on the World Wide Web, one usually performs a query through a term-based search engine. These engines return, as the query's result, a list of Web sites whose contents match the query. For broad topic queries, such searches often result in a huge set of retrieved documents, many of which are irrelevant to the user. However, much information is contained in the link-structure of the World Wide Web. Information such as which pages are linked to others can be used to augment search algorithms. In this context, Jon Kleinberg introduced the notion of two distinct types of Web sites: *hubs* and *authorities*. Kleinberg argued that hubs and authorities exhibit a *mutually reinforcing relationship*: a good hub will point to many authorities, and a good authority will be pointed at by many hubs. In light of this, he devised an algorithm aimed at finding authoritative sites. We present SALSA, a new stochastic approach for link structure analysis, which examines random walks on graphs derived from the link structure. We show that both SALSA and Kleinberg's mutual reinforcement approach employ the same meta-algorithm. We then prove that SALSA is equivalent to a weighted in-degree analysis of the link-structure of World Wide Web subgraphs, making it computationally more efficient than the mutual reinforcement approach. We compare the results of applying SALSA to the results derived through Kleinberg's approach. These comparisons reveal a topological phenomenon called the *TKC effect* (Tightly Knit Community) which, in certain cases, prevents the mutual reinforcement approach from identifying meaningful authorities.
© 2000 Published by Elsevier Science B.V. All rights reserved.

Keywords: Information retrieval; Link structure analysis; Hubs and authorities; Random walks; SALSA

1. Introduction

Searching the World Wide Web — the challenge. The World Wide Web is a rapidly expanding hyperlinked collection of unstructured information. The lack of structure and the enormous volume of the World Wide Web pose tremendous challenges on the World Wide Web information retrieval systems called search engines. These search engines are presented with queries, and return a list of Web sites which are deemed (by the engine) to pertain to the query.

When considering the difficulties which World Wide Web search engines face, we distinguish between narrow-topic queries and broad-topic queries. This distinction pertains to the presence which the query's topic has on the Web. Narrow topic queries are queries for which very few resources exist on the Web, and which present a 'needle in the haystack' challenge for search engines. An example for such a

* Corresponding author. E-mail: {rlempel, moran}@cs.technion.ac.il

[1] Abridged version

1389-1268/00/$ – see front matter © 2000 Published by Elsevier Science B.V. All rights reserved.
PII: S1389-1268(00)00034-7

query is an attempt to locate the lyrics of a specific song, by quoting a line from it ('We all live in a yellow submarine'). Search engines encounter a *recall* challenge when handling such queries: finding the few resources which pertain to the query.

On the other hand, broad-topic queries pertain to topics for which there is an abundance of information on the Web, sometimes as many as millions of relevant resources (with varying degrees of relevance). The vast majority of users are not interested in retrieving the entire huge set of resources. Most users will be quite satisfied with a few *authoritative* results: Web sites which are highly relevant to the topic of the query, significantly more than most other sites. The challenge which search engines face here is one of *precision*: retrieving only the most relevant resources to the query.

This work focuses on finding authoritative resources which pertain to broad-topic queries.

Term-based search engines. Term-based search engines face both classical problems in information retrieval, as well as problems specific to the World Wide Web setting, when handling broad-topic queries. The classic problems include the following issues [4,20].

- Synonymy — retrieving documents containing the term 'car' when given the query 'automobile'.
- Polysemy/ambiguity — when given the query 'Jordan', should the engine retrieve pages pertaining to the Hashemite Kingdom of Jordan, or pages pertaining to basketball legend Michael Jordan?
- Authorship styles — this is a generalization of the synonymy issue. Two documents, which pertain to the same topic, can sometimes use very different vocabularies and figures of speech when written by different authors (as an example, the styles of two documents, one written in British English and the other in American English, might differ considerably).

In addition to the classical issues in information retrieval, there is a Web-specific obstacle which search engines must overcome, called *search engine persuasion* [19]. There may be millions of sites pertaining in some manner to broad-topic queries, but most users will only browse through the first ten results returned by their favorite search facility. With the growing economic impact of the World Wide Web, and the growth of e-commerce, it is crucial for businesses to have their sites ranked high by the major search engines. There are quite a few companies who sell this kind of expertise. They design Web sites which are tailored to rank high with specific queries on the major search engines. These companies research the ranking algorithms and heuristics of term-based engines, and know how many keywords to place (and where) in a Web page so as to improve the page's ranking (which directly impacts the page's visibility). A less sophisticated technique, used by some site creators, is called *keyword spamming* [4]. Here, the authors repeat certain terms (some of which are only remotely connected to their site's context), in order to 'lure' search engines into ranking them highly for many queries.

Informative link structure — the answer? The World Wide Web is a hyperlinked collection. In addition to the textual content of the individual pages, the link structure of such collections contains information which can, and should, be tapped when searching for authoritative sources. Consider the significance of a link $p \to q$: with such a link p suggests, or even recommends, that surfers visiting p follow the link and visit q. This may reflect the fact that pages p and q share a common topic of interest, and that the author of p thinks highly of q's contents. Such a link, called an *informative link*, is p's way to confer authority on q [16]. Note that informative links provide a positive critical assessment of q's contents which originates from outside the control of the author of q (as opposed to assessments based on q's textual content, which is under complete control of q's author). This makes the information extracted from informative links less vulnerable to manipulative techniques such as spamming.

Unfortunately, not all links are informative. There are many kinds of links which confer little or no authority [4], such as intra-domain (inner) links (whose purpose is to provide navigational aid in a complex Web site of some organization), commercial/sponsor links, and links which result from link-exchange agreements. A crucial task which should be completed prior to analyzing the link structure of a given collection, is to filter out as many of the non-informative links as possible.

Related work on link structures. Prior to the World Wide Web age, link structures were studied in the area of bibliometrics, which studies the citation structure of written documents [15,23]. Many works in this area were aimed at finding high-impact papers published in scientific journals [10], and at clustering related documents [1].

Some works have studied the Web's link structure, in addition to the textual content of the pages, as means to visualize areas thought to contain good resources [3]. Other works used link structures for categorizing pages and clustering them [21,24].

Marchiori [19] uses the link-structure of the Web to enhance search results of term-based search engines. This is done by considering the potential hyper-information contained in each Web page: the information that can be found when following hyperlinks which originate in the page.

This work is motivated by the approach introduced by Jon Kleinberg [16]. In an attempt to impose some structure on the chaotic World Wide Web, Kleinberg distinguished between two types of Web sites which pertain to a certain topic. The first are *authoritative* pages in the sense described previously. The second type of sites are *hub* pages. Hubs are resource lists. They do not directly contain information pertaining to the topic, but rather point to many authoritative sites. According to this model, hubs and authorities exhibit a *mutually reinforcing relationship*: good hubs point to many good authorities, and good authorities are pointed at by many good hubs.

In light of the mutually reinforcing relationship, hubs and authorities should form communities, which can be pictured as dense bipartite portions of the Web, where the hubs link densely to the authorities. The most prominent community in a World Wide Web subgraph is called the *principal community* of the collection. Kleinberg suggested an algorithm to identify these communities, which is described in detail in Section 2.

Researchers from IBM's Almaden Research Center have implemented Kleinberg's algorithm in various projects. The first was *HITS*, which is described in [11], and offers some enlightening practical remarks. The *ARC* system, described in [7], augments Kleinberg's link-structure analysis by considering also the anchor text, the text which surrounds the hyperlink in the pointing page. The reasoning behind this is that many times, the pointing page describes the destination page's contents around the hyperlink, and thus the authority conferred by the links can be better assessed. These projects were extended by the *CLEVER* project [14]. Researchers from outside IBM, such as Henzinger and Brahat, have also studied Kleinberg's approach and have proposed improvements to it [13].

Anchor text has also been used by Brin and Page in [2]. Another major feature of their work on the *Google* search engine [12] is a link-structure based site ranking approach called *PageRank*, which can be interpreted as a stochastic analysis of some random-walk behavior through the entire World Wide Web.

In [18], the authors use the links surrounding a small set of same-topic sites to assemble a larger collection of neighboring pages which should contain many authoritative resources on the initial topic. The textual content of the collection is then analyzed in ranking the relevancy of its individual pages.

This work. While preserving the theme that Web sites pertaining to a given topic should be split to hubs and authorities, we replace Kleinberg's mutual reinforcement approach [16] by a new stochastic approach (SALSA), in which the coupling between hubs and authorities is less tight. The intuition behind our approach is the following. Consider a bipartite graph G, whose two parts correspond to hubs and authorities, where an edge between hub r and authority s means that there is an informative link from r to s. Then, authorities and hubs pertaining to the dominant topic of the sites in G should be highly visible (reachable) from many sites in G. Thus, we will attempt to identify these sites by examining certain random walks in G, under the proviso that such random walks will tend to visit these highly visible sites more frequently than other, less connected sites. We show that in finding the principal communities of hubs and authorities, both Kleinberg's mutual reinforcement approach and our stochastic approach employ the same meta-algorithm on different representations of the input graph. We then compare the results of applying SALSA to the results derived by Kleinberg's approach. Through these comparisons, we isolate a particular topological phenomenon which we call the *Tightly Knit Community (TKC) effect*. In certain scenarios, this effect hampers the ability of the mutual reinforcement approach to identify meaningful authorities. We demonstrate that SALSA is less vulnerable to the TKC effect, and can find meaningful authorities in collections where the mutual reinforcement approach fails to do so.

After demonstrating some results achieved by means of SALSA, we prove that the ranking of sites in the stochastic approach may be calculated by examining the weighted in/out degrees of the sites in G. This result yields that SALSA is computationally lighter than the mutual reinforcement approach. We also discuss the reason for our success with analyzing weighted in/out degrees of sites, which previous work has claimed to be unsatisfactory for identifying authoritative sites.

The rest of the paper is organized as follows. Section 2 recounts Kleinberg's mutual reinforcement approach. In Section 3 we view Kleinberg's approach from a higher level, and define a meta-algorithm for link structure analysis. Section 4 presents our new approach, SALSA. In Section 5 we compare the two approaches by considering their outputs on the World Wide Web and on artificial topologies. Then, in Section 6 we prove the connection between SALSA and weighted in/out degree rankings of sites. Our conclusions and ideas for future work are brought in Section 7. The paper uses basic results from the theory of stochastic processes, which are brought in the full version. The main contribution of the paper can be grasped without following the full mathematical analysis.

2. Kleinberg's mutual reinforcement approach

The mutual reinforcement approach [16] starts by assembling a collection $\mathcal{C}$ of Web sites, which should contain communities of hubs and authorities pertaining to a given topic t. It then analyzes the link structure induced by that collection, in order to find the authoritative sites on topic t.

Denote by q a term-based search query to which sites in our topic of interest t are deemed to be relevant. The collection $\mathcal{C}$ is assembled in the following manner.

- A *root set* S of sites is obtained by applying a term-based search engine, such as AltaVista [8], to the query q. This is the only step in which the lexical content of the Web sites is examined.
- From S we derive a *base set* $\mathcal{C}$ which consists of (a) sites in the root set S, (b) sites which point to a site in S, and (c) sites which are pointed to by a site in S. In order to obtain (b), we must again use a search engine. Many search engines store linkage information, and support queries such as 'which sites point to [a given URL]'.

The collection $\mathcal{C}$ and its link structure induce the following directed graph G: G's nodes are the sites in $\mathcal{C}$, and for all $i, j \in \mathcal{C}$ the directed edge $i \to j$ appears in G if and only if site i contains a hyperlink to site j. Let W denote the $|\mathcal{C}| \times |\mathcal{C}|$ adjacency matrix of G.

Each site $s \in \mathcal{C}$ is now assigned a pair of weights, a hub weight $h(s)$ and an authority weight $a(s)$, based on the following two principles:

- The quality of a hub is determined by the quality of the authorities it points at. Specifically, a site's hub weight should be proportional to the sum of the authority weights of the sites it points at.
- 'Authority lies in the eyes of the beholder(s)': a site is authoritative only if good hubs deem it as such. Hence, a site's authority weight is proportional to the sum of the hub weights of the sites pointing at it.

The top ranking sites, according to both kinds of weights, form the mutually reinforcing communities of hubs and authorities. In order to assign such weights, Kleinberg uses the following iterative algorithm:

(1) Initialize $a(s) \leftarrow 1$, $h(s) \leftarrow 1$ for all sites $s \in \mathcal{C}$.
(2) Repeat the following three operations until convergence:
   - Update the authority weight of each site s (the $\mathcal{I}$ operation):
   $$a(s) \leftarrow \sum_{x \mid x \text{ points to } s} h(x)$$
   - Update the hub weight of each site s (the $\mathcal{O}$ operation):
   $$h(s) \leftarrow \sum_{x \mid s \text{ points to } x} a(x)$$
   - Normalize the authority weights and the hub weights.

Note that applying the $\mathcal{I}$ operation is equivalent to assigning authority weights according to the result of multiplying the vector of all hub weights by the matrix W^T. The $\mathcal{O}$ operation is equivalent to assigning hub weights according to the result of multiplying the vector of all authority weights by the matrix W.

Kleinberg showed that this algorithm converges, and that the resulting authority weights [hub weights] are the coordinates of the normalized principal eigenvector[2] of W^TW [of WW^T]. W^TW and WW^T are well known matrices in the field of bibliometrics:

(1) $A \overset{\triangle}{=} W^TW$ is the *co-citation matrix* [23] of the collection. $[A]_{i,j}$ is the number of sites which jointly point at (cite) pages i and j. Kleinberg's iterative algorithm converges to authority weights which correspond to the entries of the (unique, normalized) principal eigenvector of A.

(2) $H \overset{\triangle}{=} WW^T$ is the *bibliographic coupling matrix* [15] of the collection. $[H]_{i,j}$ is the number of sites jointly referred to (pointed at) by pages i and j. Kleinberg's iterative algorithm converges to hub weights which correspond to the entries of H's (unique, normalized) principal eigenvector.

3. A meta-algorithm for link structure analysis

Examining the mutual reinforcement approach from a higher level, we can identify a general framework, or meta-algorithm, for finding hubs and authorities by link structure analysis. This meta-algorithm is a version of the spectral filtering method, presented in [6].

- Given a topic t, construct a site collection $\mathcal{C}$ which should contain many t-hubs and t-authorities, but should not contain many hubs or authorities for any other topic t'. Let $n = |\mathcal{C}|$.
- Derive, from $\mathcal{C}$ and the link structure induced by it, two $n \times n$ association matrices: a *hub matrix H* and an *authority matrix A*. Association matrices are widely used in classification algorithms [22] and will be used here in order to classify the Web sites into communities of hubs/authorities. The association matrices which are used by the meta-algorithm will have the following algebraic property (let M denote such a matrix). M will have a unique real positive eigenvalue $\mu(M)$ of multiplicity 1, such that for any other eigenvalue μ' of M, $\mu(M) > |\mu'(M)|$. Denote by $v_{\mu(M)}$ the (unique) unit eigenvector which corresponds to $\mu(M)$ whose first non-zero coordinate is positive. $v_{\mu(M)}$ will actually be a positive vector, and will be referred to as the *principal eigenvector* of M.
- The sites that correspond to the largest coordinates of $v_{\mu(A)}$ will form the *principal algebraic community of authorities* in $\mathcal{C}$, and the sites that correspond to the largest coordinates of $v_{\mu(H)}$ will form the *principal algebraic community of hubs* in $\mathcal{C}$.

For the meta-algorithm to be useful, the algebraic principal communities of hubs and authorities should reflect the true authorities and hubs in $\mathcal{C}$.

The two degrees of freedom which the meta-algorithm allows, are the method for obtaining the collection, and the definition of the association matrices. Given a specific collection, the algebraic communities produced by the meta-algorithm are determined solely by the definition of the association matrices.

4. SALSA: analyzing a random walk on the Web

In this section we introduce the *stochastic approach for link structure analysis* (SALSA). The approach is based upon the theory of Markov chains, and relies on the stochastic properties of random walks performed on our collection of sites. It follows the meta-algorithm described in Section 3, and differs from the mutual reinforcement approach in the manner in which the association matrices are defined.

The input to our scheme consists of a collection of sites $\mathcal{C}$ which is built around a topic t in the manner described in Section 2. Intuition suggests that authoritative sites on topic t should be visible from many sites in the subgraph induced by scC. Thus, a random walk on this subgraph will visit t-authorities with high probability.

We combine the theory of random walks with the notion of the two distinct types of Web sites, hubs and authorities, and actually analyze two different Markov chains: a chain of hubs and a chain of authorities. Unlike 'conventional' random walks on graphs, state transitions in these chains are generated by traversing *two* World Wide Web links in a row,

[2] The eigenvector which corresponds to the eigenvalue of highest magnitude of the matrix.

one link forward and one link backwards (or vice versa). Analyzing both chains allows our approach to give each Web site two distinct scores, a hub score and an authority score.

The idea of ranking Web sites using random walks is not new. The search engine *Google* [2,12] incorporates stochastic information into its ranking of pages. The *PageRank* component of the search engine examines a *single* random walk on the *entire* World Wide Web. Hence, the ranking of Web sites in *Google* is independent of the search query (a global ranking), and no distinction is made between hubs and authorities.

Let us build a bipartite undirected graph $\tilde{G} = (V_h, V_a, E)$ from our site collection $\mathcal{C}$ and its link structure:

- $V_h = s_h | s \in \mathcal{C}$ and *out-degree(s)* >0 (the *hub side* of $\tilde{G}$).
- $V_a = s_a | s \in \mathcal{C}$ and *in-degree(s)* >0 (the *authority side* of $\tilde{G}$).
- $E = (s_h, r_a) | s \to r$ in $\mathcal{C}$

Each non-isolated site $s \in \mathcal{C}$ is represented by two nodes of $\tilde{G}$, s_h and s_a. Each World Wide Web link $s \to r$ is represented by an undirected edge connecting s_h and r_a.

On this bipartite graph we will perform two distinct random walks. Each walk will only visit nodes from one of the two sides of the graph, by traversing paths consisting of two $\tilde{G}$-edges in each step. Since each edge crosses sides of $\tilde{G}$, each walk is confined to just one of the graph's sides, and the two walks will naturally start off from different sides of $\tilde{G}$. Note also that every path of length 2 in $\tilde{G}$ represents a traversal of one World Wide Web link in the proper direction (when passing from the hub side of $\tilde{G}$ to the authority side), and a retreat along a World Wide Web link (when crossing in the other direction). Since the hubs and authorities of topic t should be highly visible in $\tilde{G}$ (reachable from many nodes by either a direct edge or by short paths), we may expect that the t-authorities will be amongst the nodes most frequently visited by the random walk on V_a, and that the t-hubs will be amongst the nodes most frequently visited by the random walk on V_h.

We will examine the two different Markov chains which correspond to these random walks: the chain of the visits to the authority side of $\tilde{G}$ (the *authority chain*), and the chain of visits to the hub side of $\tilde{G}$. Analyzing these chains separately naturally distinguishes between the two aspects of each site.

We now define two stochastic matrices, which are the transition matrices of the two Markov chains at interest.

(1) *The hub matrix* $\tilde{H}$, defined as follows:

$$\tilde{h}_{i,j} = \sum_{k|(i_h,k_a),(j_h,k_a)\in\tilde{G}} \frac{1}{\deg(i_h)} \times \frac{1}{\deg(k_a)}$$

(2) *The authority matrix* $\tilde{A}$, defined as follows:

$$\tilde{a}_{i,j} = \sum_{k|(k_h,i_a),(k_h,j_a)\in\tilde{G}} \frac{1}{\deg(i_a)} \times \frac{1}{\deg(k_h)}$$

A positive transition probability $\tilde{a}_{i,j} > 0$ implies that a certain page h points to both pages i and j, and hence page j is reachable from page i by two steps: retracting along the link $h \to i$ and then following the link $h \to j$.

Alternatively, the matrices $\tilde{H}$ and $\tilde{A}$ can be defined as follows. Let W be the adjacency matrix of the directed graph defined by $\mathcal{C}$ and its link structure. Denote by W_r the matrix which results by dividing each non-zero entry of W by the sum of the entries in its row, and by W_c the matrix which results by dividing each non-zero element of W by the sum of the entries in its column. (Obviously, the sums of rows/columns which contain non-zero elements are greater than zero.) Then $\tilde{H}$ consists of the non-zero rows and columns of $W_r W_c^{\mathrm{T}}$, and $\tilde{A}$ consists of the non-zero rows and columns of $W_c^{\mathrm{T}} W_r$. We ignore the rows and columns of $\tilde{A}$, $\tilde{H}$ which consist entirely of zeros, since (by definition) all the nodes of $\tilde{G}$ have at least one incident edge. The matrices $\tilde{A}$ and $\tilde{H}$ serve as the association matrices required by the meta-algorithm for identifying the authorities and hubs. Recall that the mutual reinforcement approach uses the association matrices $A \triangleq W^{\mathrm{T}} W$ and $H \triangleq W W^{\mathrm{T}}$.

We shall assume that $\tilde{G}$ is connected, causing both stochastic matrices $\tilde{A}$ and $\tilde{H}$ to be *irreducible.* This assumption does not form a limiting factor, since when $\tilde{G}$ is not connected, we may use our technique on each connected component separately. Section 6.1 further elaborates on the case when $\tilde{A}$ and $\tilde{H}$ have multiple irreducible components.

Some properties of $\tilde{H}$ and $\tilde{A}$:

- Both matrices are primitive, since the Markov chains which they represent are aperiodic: when visiting any authority (hub), there is a positive probability to revisit it on the next entry to the authority (hub) side of the bipartite graph (since all the nodes are non-isolated). Hence, every state (= site) in each of the chains has a self-loop, causing the chains to be aperiodic.

- The adjacency matrix of the support graph of $\tilde{A}$ is symmetric, since $\tilde{a}_{i,j} > 0$implies $\tilde{a}_{j,i} > 0$. Furthermore, $\tilde{a}_{i,j} > 0 \Leftrightarrow [W^{\mathrm{T}}W]_{i,j} > 0$ (and the same is also true of $\tilde{H}$ and WW^{T}).

Following the framework of the meta-algorithm, the principal community of authorities (hubs) found by the SALSA will be composed of the sites whose entries in the principal eigenvector of $\tilde{A}$ ($\tilde{H}$) are the highest. By the ergodic theorem [9], the principal eigenvector of an irreducible, aperiodic stochastic matrix is actually the stationary distribution of the underlying Markov chain, and its high entries correspond to sites most frequently visited by the (infinite) random walk.

5. Results

5.1. The tightly knit community (TKC) effect

A tightly knit community is a small but highly interconnected set of sites. Roughly speaking, the TKC effect occurs when such a community scores high in link-analyzing algorithms, even though the sites in the TKC are not authoritative on the topic, or pertain to just one aspect of the topic. Our study indicates that the mutual reinforcement approach is vulnerable to this effect, and will sometimes rank the sites of a TKC in unjustified high positions.

As an example, consider a collection $\mathcal{C}$ which contains the following two communities: a community y, with a small number of hubs and authorities, in which every hub points to most of the authorities, and a much larger community z, in which each hub points to a smaller part of the authorities. The topic covered by z is the dominant topic of the collection, and is probably of wider interest on the World Wide Web. Since there are many z-authoritative sites, the hubs do not link to all of them, whereas the smaller y community is densely interconnected. The TKC effect occurs when the sites of y are ranked higher than those of z.

In the full paper we provide a combinatorial construction, which demonstrates such (artificial) communities y and z, where the mutual reinforcement approach scores y higher than z, and the stochastic approach scores z higher. This bias of the mutual reinforcement approach towards tightly knit communities will be demonstrated on World Wide Web queries in the next section.

5.2. The World Wide Web

We tested the different approaches on broad-topic World Wide Web queries (both single-topic queries and multi-topic queries). We obtained a collection of sites for each query, and then derived the principal community of authorities with both approaches. Two of these queries (‘Java’, ‘abortion’) were used by Kleinberg in [16], and are brought here for the sake of comparison. All collections were assembled during February, 1999. The root sets were compiled using AltaVista [8], which also provided the linkage information needed for building the base sets.

When expanding the root set to the entire collection, we filtered the links pointing to and from Web sites. Following [16], we ignored intra-domain links (since these links tend to be navigational aids inside an intranet, and do not confer authority on the link’s destination). We also ignored links to *cgi scripts*, and tried to identify ad-links and ignore them as well. Overall, 38% of the links we examined were ignored. The collections themselves turn out to be relatively sparse graphs, with the number of edges never exceeding three times the number of nodes. We note that a recent work by Kleinberg et al. [17] has examined some other connectivity characteristics of such collections.

For each query, we list the top authorities which were returned by the two approaches. The results are displayed in tables containing four columns:

(1) The URL.
(2) The title of the URL.
(3) The category of the URL: (a) for a member of the root set, (b) for a site pointing into the root set, and (c) for a site pointed at by a member of the root set.
(4) The value of the coordinate of this URL in the principal eigenvector of the authority matrix.

Table 1
Authorities for World Wide Web query 'Java' (size of root size = 160, size of collection = 2810)

URL	Title	Cat.	Weight
Principal community, mutual reinforcement approach			
http://www.jars.com/	EarthWeb's JARS.COM Java Review Service	(3)	0.334102
http://www.gamelan.com/	Gamelan — The Official Java Directory	(3)	0.303624
http://www.javascripts.com/	Javascripts.com — Welcome	(3)	0.255254
http://www.datamation.com/	EarthWeb's Datamation.com	(3)	0.251379
http://www.roadcoders.com/	Handheld Software Development@RoadCoders	(3)	0.250816
http://www.earthweb.com/	EarthWeb	(3)	0.249373
http://www.earthwebdirect.com/	Welcome to Earthweb Direct	(3)	0.247467
http://www.itknowledge.com/	ITKnowledge	(3)	0.246874
http://www.intranetjournal.com/	intranetjournal.com	(3)	0.24518
http://www.javagoodies.com/	Java Goodies JavaScript Repository	(3)	0.238793
Principal community, SALSA			
http://java.sun.com/	Java(tm) Technology Home Page	(3)	0.365264
http://www.gamelan.com/	Gamelan — The Official Java Directory	(3)	0.36369
http://www.jars.com/	EarthWeb's JARS.COM Java Review Service	(3)	0.303862
http://www.javaworld.com/	IDG's magazine for the Java community	(3)	0.217269
http://www.yahoo.com/	Yahoo!	(3)	0.21412
http://www.javasoft.com/	Java(tm) Technology Home Page	(3)	0.203099
http://www.sun.com/	Sun Microsystems	(3)	0.187355
http://www.javascripts.com/	Javascripts.com — Welcome	(3)	0.138548
http://www.htmlgoodies.com/	htmlgoodies.com — Home	(3)	0.130676
http://javaboutique.internet.com/	The Ultimate Java Applet Resource	(1)	0.118081

5.2.1. Single-topic query: Java

The results for this query, with our first example of the TKC effect, are shown in Table 1. All of the top ten mutual reinforcement authorities are part of the EarthWeb Inc. network. They are interconnected, but since the domain names of the sites are different, the interconnecting links were not filtered out. Some of the sites are highly relevant to the query (and have many incoming links from sites outside the EarthWeb net), but most appear in the principal community only because of their EarthWeb affiliation. With SALSA, only the top three mutual reinforcement authorities are retained, and the other seven are replaced by other authorities, some of which are clearly more related to the query.

5.2.2. Single-topic query: movies

This query demonstrates the TKC effect in a most striking fashion on the World Wide Web. First, consider the mutual reinforcement principal community of authorities, presented in Table 2.

The top 30 authorities returned by the mutual reinforcement approach were all *go.msn.com* sites. All but the first received the exact same weight, 0.167202. Recall that we do not allow same-domain links in our collection, hence none of the top authorities was pointed at by a *go.msn.com* site. To understand how these sites scored so well, we turn to the principal community of hubs, shown in Table 3.

These innocent looking hubs are all part of the *Microsoft Network (msn)*, but when building the ba-

Table 2
Mutual Reinforcement Authorities for World Wide Web query 'movies' (size of root size = 175, size of collection = 4539)

URL	Title	Cat	Weight
http://go.msn.com/npl/msnt.asp	MSN.COM	(3)	0.167332
http://go.msn.com/bql/whitepages.asp	White Pages — msn.com	(3)	0.167202
http://go.msn.com/bsl/webevents.asp	Web Events	(3)	0.167202
http://go.msn.com/bql/scoreboards.asp	MSN Sports scores	(3)	0.167202

Table 3
Mutual reinforcement hubs for World Wide Web query ‘movies’

URL	Title	Cat	Weight
http://denver.sidewalk.com/movies	movies: denver.sidewalk	(1)	0.169197
http://boston.sidewalk.com/movies	movies:boston.sidewalk	(1)	0.169061
http://twincities.sidewalk.com/movies	movies: twincities.sidewalk	(1)	0.1688
http://newyork.sidewalk.com/movies	movies: newyork.sidewalk	(1)	0.168537

sic set we did not identify them as such. All these hubs point, almost without exception, to the entire set of authorities found by the MR approach (hence the equal weights which the authorities exhibit). However, the vast majority of the sites in the collection were not part of this ‘conspiracy’, and almost never pointed to any of the *go.msn.com* sites. Therefore, the authorities returned by the stochastic approach (Table 4) contain none of those *go.msn.com* sites, and are much more relevant to the query.

A similar community is obtained by the mutual reinforcement approach, after deleting the rows and columns which correspond to the top 30 authorities from the matrix $W^{\mathrm{T}}W$. This deletion dissolves the *msn.com* community, and allows a community similar to the one obtained by SALSA to manifest itself.

5.2.3. Multi-topic query: abortion

This topic is highly polarized, with different cyber communities supporting pro-life and pro-choice views. In Table 5, we bring the top 10 authorities, as determined by the two approaches.

All 10 top authorities found by the mutual reinforcement approach are pro-life resources, while the top 10 SALSA authorities are split, with 6 pro-choice sites and 4 pro-life sites (which are the same top 4 pro-life sites found by the mutual reinforcement approach). Again, we see the TKC effect: the mutual reinforcement approach ranks highly authorities on only one aspect of the query, while SALSA blends authorities from both aspects into its principal community.

5.2.4. Multi-topic query: genetics

This query is especially ambiguous in the World Wide Web: it can be in the context of genetic engineering, genetic algorithms, or in the context of health issues and the human genome.

As in the ‘abortion’ query, SALSA brings a diverse principal community, with authorities on the various contexts of the query, while the mutual reinforcement approach is focussed on one context (genetic algorithms, in this case). Both principal communities are shown in Table 6.

6. SALSA and the in/out degrees of sites

In the previous sections we have presented the stochastic approach as an alternative method for link-structure analysis, and have shown a few encouraging results obtained by it, as compared with the mutual reinforcement approach. We have also presented the TKC effect, a topological phenomenon which sometimes derails the MR approach and prevents it from converging to a useful community of authoritative sites.

Table 4
Stochastic authorities for World Wide Web query ‘movies’

URL	Title	Cat	Weight
http://us.imdb.com/	The Internet Movie Database	(3)	0.253333
http://www.mrshowbiz.com/	Mr Showbiz	(3)	0.22335
http://www.disney.com/	Disney.com — The Web Site for Families	(3)	0.22003
http://www.hollywood.com/	Hollywood Online: ...all about movies	(3)	0.213355
http://www.imdb.com/	The Internet Movie Database	(3)	0.199987
http://www.paramount.com/	Welcome to Paramount Pictures	(3)	0.196682
http://www.mca.com/	Universal Studios	(3)	0.180021

Table 5
Authorities for World Wide Web query 'abortion' (size of root size = 160, size of collection = 1693)

URL	Title	Cat	Weight
Principal community, mutual reinforcement approach			
http://www.nrlc.org/	National Right To Life	(3)	0.420832
http://www.prolife.org/ultimate/	The Ultimate Pro-Life Resource List	(3)	0.316564
http://www.all.org/	What's new at American Life League	(3)	0.251506
http://www.hli.org/	Human Life International	(3)	0.212931
http://www.prolife.org/cpcs-online/	Crisis Pregnancy Centers Online	(3)	0.187707
http://www.ohiolife.org/	Ohio Right to Life	(3)	0.182076
http://www.rtl.org/	Abortion, adoption assisted-suicide, Information at Right to Life...	(1)	0.17943
http://www.bethany.org/	Bethany Christian Services	(3)	0.161359
http://www.ldi.org/	Abortion malpractice litigation	(1)	0.140076
http://www.serve.com/fem4life/	Feminists for Life of America	(3)	0.122106
Principal community, SALSA			
http://www.nrlc.org/	National Right To Life	(3)	0.344029
http://www.prolife.org/ultimate/	The Ultimate Pro-Life Resource List	(3)	0.284714
http://www.naral.org/	NARAL Choice for America	(3)	0.240227
http://www.feminist.org/	Feminist Majority Foundation	(3)	0.186843
http://www.now.org/	National Organization for Women	(3)	0.177946
http://www.cais.com/agm/main/index.html	The Abortion Rights Activist	(1)	0.166083
http://www.gynpages.com/	Abortion Clinics Online	(3)	0.163117
http://www.plannedparenthood.org/	Planned Parenthood Federation	(3)	0.157186
http://www.all.org/	What's new at American Life League	(3)	0.142357
http://www.hli.org/	Human Life International	(3)	0.142357

Table 6
Authorities for World Wide Web query 'genetic' (size of root size = 120, size of collection = 2952

URL	Title	Cat	Weight
Principal community, mutual reinforcement approach			
http://www.aic.nrl.navy.mil/galist/	The Genetic Algorithms Archive	(3)	0.27848
http://alife.santafe.edu/	Artificial Life Online	(3)	0.276159
http://www.yahoo.com/	Yahoo!	(3)	0.273599
http://www.geneticprogramming.com/	The Genetic Programming Notebook	(1)	0.25588
http://gal4.ge.uiuc.edu/illigal.home.html	illiGAL Home Page	(3)	0.235717
http://www.cs.gmu.edu/research/gag/	The Genetic Algorithms Group...	(3)	0.201237
http://www.scs.carleton.ca/ csgs/resources/gaal.html	Genetic Algorithms and Artificial Life Resources	(1)	0.181315
http://lancet.mit.edu/ga/	GAlib: Matthew's Genetic Algorithms Library	(3)	0.181157
Principal community, SALSA			
http://www.ncbi.nlm.nih.gov/	The National Center for Biotechnology Information	(3)	0.250012
http://www.yahoo.com/	Yahoo!	(3)	0.227782
http://www.aic.nrl.navy.mil/galist/	The Genetic Algorithms Archive	(3)	0.223191
http://www.nih.gov/	National Institute of Health (NIH)	(3)	0.194688
http://gdbwww.gdb.org/	The Genome Database	(3)	0.177001
http://alife.santafe.edu/	Artificial Life Online	(3)	0.172383
http://www.genengnews.com/	Genetic Engineering News (GEN)	(1)	0.141617
http://gal4.ge.uiuc.edu/illigal.home.html	illiGAL Home Page	(3)	0.13259

The sample results shown so far have all been produced on unweighted collections, in which all informative links have received unit weight. Both approaches can produce better rankings when applied on weighted collections, in which each informative link receives a weight which reflects the amount of authority that the pointing site confers to the pointed site. Possible factors which may contribute to a link's weight include the following.

- Anchor text which is relevant to the query. Such text around a link heightens our confidence that the pointed site discusses the topic at hand [7].
- One of the link's endpoints being designated by the user as highly relevant to the search topic. When a site points to one of a small set of predefined authorities, it seems reasonable to raise the weights of other links which originate from that site. Similarly, when a site is known to be a good hub, it seems reasonable to assign high weights to its outgoing links. This approach has been recently applied in [5]. We coin it the *anchor sites* approach, since it uses user-designated sites as anchors in the collection, around which the communities of hubs and authorities are grown.
- The link's placement in the pointing page. Many search engines consider the text at the top of a page as more reflective of its contents than text further down the page. The same line of thought can be applied to the links which appear in a page, with the links which are closer to the top of the page receiving more weight than links appearing at the bottom of the page.

6.1. Analysis of the stochastic ranking

We now prove a general result about the ranking produced by SALSA in weighted collections, for which some basic background in stochastic processes is assumed.

Let $G = (H; A; E)$ be a positively weighted, directed bipartite graph with no isolated nodes, and let all edges be directed from sites in H to sites in A. We will use the following notations:

- The weighted in-degree of site $i \in A$:

$$d_{\text{in}}(i) \stackrel{\triangle}{=} \sum_{k \in H | k \rightarrow i} w(k \rightarrow i)$$

- The weighted out-degree of site $k \in H$:

$$d_{\text{out}}(k) \stackrel{\triangle}{=} \sum_{i \in A | k \rightarrow i} w(k \rightarrow i)$$

- The sum of edge weights:

$$\mathcal{W} = \sum_{i \in A} d_{\text{in}}(i) = \sum_{k \in H} d_{\text{out}}(k)$$

Let M_A be a Markov chain whose states are the set A of vertices, with the following transition probabilities between every two states $i, j \in A$:

$$P_A(i, j) = \sum_{k \in H | k \rightarrow i, k \rightarrow i} \frac{w(k \rightarrow i)}{d_{\text{in}}(i)} \times \frac{w(k \rightarrow j)}{d_{\text{out}}(k)}$$

Similarly, let M_H be a Markov chain whose states are the set H of vertices, with the following transition probabilities between every two states $k, l \in H$:

$$P_H(k, l) = \sum_{i \in A | k \rightarrow i, l \rightarrow i)} \frac{w(k \rightarrow i)}{d_{\text{out}}(k)} \times \frac{w(l \rightarrow i)}{d_{\text{in}}(i)}$$

Consider the following binary relation on the vertices of A (states of M_A):

$$R_A = (i, j) | P_A(i, j) > 0$$

It is not hard to show (and is shown in the full paper) that R_A is an equivalence relation on A (similar arguments can be made concerning M_H). This implies that all the states of M_A are recurrent (none are transient). The equivalence classes of R_A are the irreducible components of M_A. We first deal with the case where R_A consists of one equivalence class (i.e., M_A is irreducible).

Proposition 1. *Whenever M_A is an irreducible chain (has a single irreducible component), it has a unique stationary distribution $\pi = (\pi_1, \ldots, \pi_{|A|})$ satisfying:*

$$\pi_i = \frac{d_{\text{in}}(i)}{\mathcal{W}} \quad \textit{for all } i \in A$$

Similarly, whenever M_H is an irreducible chain, its unique stationary distribution $\pi = (\pi_1, \ldots, \pi_{|H|})$ satisfies:

$$\pi_k = \frac{d_{\text{out}}(k)}{\mathcal{W}} \quad \textit{for all } k \in H$$

Proof. We will prove the proposition for M_A. The proof for M_H is similar.

By the ergodic theorem [9], any irreducible, aperiodic Markov chain has a unique stationary distribution vector. It will therefore suffice to show that the vector π with the properties claimed in the proposition is indeed a stationary distribution vector of M_A.

- π is a distribution vector: its entries are non-negative, and their sum equals one.

$$\sum_{i \in A} \pi_i = \sum_{i \in A} \frac{d_{\text{in}}(i)}{\mathcal{W}} = \frac{1}{\mathcal{W}} \sum_{i \in A} d_{\text{in}}(i) = 1$$

- π is a stationary distribution vector of M_A. Here we need to show the equality $\pi P_A = \pi$:

$$
\begin{aligned}
[\pi P_A]_i &= \sum_{j\in A} \pi_j P_A(j,i)\\
&= \sum_{j\in A} \frac{d_{\mathrm{in}}(j)}{\mathcal{W}} \sum_{k\in H|k\to i,k\to j} \frac{w(k\to j)}{d_{\in}(j)} \frac{w(k\to i)}{d_{\mathrm{out}}(k)}\\
&= \frac{1}{\mathcal{W}} \sum_{j\in A} \sum_{k\in H|k\to i,k\to j} \frac{w(k\to j)w(k\to i)}{d_{\mathrm{out}}(k)}\\
&= \frac{1}{\mathcal{W}} \sum_{k\in H|k\to i} \sum_{j\in A|k\to j} \frac{w(k\to j)w(k\to i)}{d_{\mathrm{out}}(k)}\\
&= \frac{1}{\mathcal{W}} \sum_{k\in H|k\to i} \frac{w(k\to i)}{d_{\mathrm{out}}(k)} \sum_{j\in A|k\to j} w(k\to j)\\
&= \frac{1}{\mathcal{W}} \sum_{k\in H|k\to i} w(k\to i)\\
&= \frac{d_{\in}(i)}{\mathcal{W}}\\
&= \pi_i \qquad \square
\end{aligned}
$$

Thus, when the (undirected) support graph of G is connected, SALSA assigns each site an authority weight which is proportional to the sum of weights of its incoming edges. The hub weight of each site is proportional to the sum of weights of its outgoing edges. In unweighted collections (with all edges having unit weight), each site's stochastic authority (hub) weight is simply proportional to the in-(out-)degree of the site.

This mathematical analysis, in addition to providing insight about the ranking that is produced by SALSA, also suggests a very simple algorithm for calculating the stochastic ranking: simply calculate, for all sites, the sum of weights on their incoming (outgoing) edges, and normalize these two vectors. There is no need to apply any resource-consuming iterative method to approximate the principal eigenvector of the transition matrix of the Markov chain.

Markov chains with multiple irreducible components. Consider the case in which the authority chain M_A consists of multiple irreducible components. Denote these (pairwise disjoint) components by $A_1, A_2, \ldots, A_k$ where $A_i \subset A$, $1 \le i \le k$. What will be the outcome of a random walk performed on the set of states A according to the transition matrix P_A? To answer this question, we will need some notations:

- Let e denote the $|A|$-dimensional distribution vector, all whose entries equal $1/|A|$.
- For all vertices $j \in A$, denote by $c(j)$ the irreducible component (equivalence class of R_A) to which J belongs: $c(j) = l \Leftrightarrow j \in A_l$.
- Let $\pi^1, \pi^2, \ldots, \pi^k$ be the unique stationary distributions of the (irreducible) Markov chains induced by $A_1, \ldots, A_k$.
- Denote by $\pi^{c(j)_j}$ the entry which corresponds to j in $\pi^{c(j)}$ (the stationary distribution of j's irreducible component, $A_{c(j)}$).

Proposition 2. *The random walk on A, governed by the transition matrix P_A and started from all states with equal probability, will converge to a stationary distribution as follows:*

$$\lim\nolimits_{n\to\infty} eP_A^n = \tilde{\pi} \quad \textit{where} \quad \tilde{\pi}_j = \frac{|A_{c(j)|}|}{|A|}\pi^{c(j)_j}$$

Proof. Denote by p_i^n, $1 \le i \le k$ the probability of being in a site belonging to A_i after the nth step of the random walk. This probability is determined by the distribution vector eP_A^n. Clearly,

$$p_i^0 = \sum_{j\in A_i} e_j = \frac{|A_i|}{|A|}$$

Since the transition probability between any two sites (states) which belong to different irreducible components is zero, $p_i^n = p_i^0$ for all n (probability does not shift from one component to another). Inside each irreducible component the ergodic theorem holds, thus the probabilities which correspond to the sites of A_i in $\lim_{n\to\infty} eP_A^n$ will be proportional to π^i, and the proposition follows. □

This proposition points out a natural way to compare the authoritativeness of sites from different irreducible components: simply multiply each site's authority score by the normalized size of the irreducible component to which it belongs. We do not claim that this is in any way optimal, as very small irreducible components should be trimmed from the graph altogether. But the underlying principle is important: consider the size of the community when evaluating the quality of the top sites in that community. The budget which the Mayor of New York City controls is much larger than that of the Mayor of Osh Kosh, Wisconsin.

It is this combination of a site's intra-community authority score and its community's size that allows the stochastic approach to blend authorities from different aspects of a multi-topic query, and which reduces its vulnerability to the TKC effect.

6.2. In-degree as a measure of authority (revisited)

Extensive research in link-structure analysis has been conducted in recent years under the premise that considering the in-degree of sites as a sole measure of their authority does not produce satisfying results. Kleinberg, as a motivation to the mutual reinforcement approach, showed some examples of the inadequacy of a simple in-degree ranking [16]. Our results in Section 5.2 seem to contradict this premise: the stochastic rankings seem quite satisfactory there, and since those collections were unweighted, the stochastic rankings are equivalent to simple in-degree counts (normalized by the size of the connected component which each site belongs to). To gain more perspective on this apparent contradiction, let us elaborate on the first stage of the meta-algorithm for link-structure analysis (from Section 3), in which the graph to be analyzed is assembled:

(1) Given a query, assemble a collection of Web sites which should contain many hubs and authorities pertaining to the query, and few hubs and authorities for any particular unrelated topic.
(2) Filter out non-informative links connecting sites in the collection.
(3) Assign weights to all non-filtered links. These weights should reflect the information conveyed by the link.

It is only after these steps that the weighted, directed graph is analyzed and the rankings of hubs and authorities are produced. The analysis of the graph, however important, is just the second stage in the meta-algorithm, and the steps involved in the first stage are crucial to the success of the entire algorithm.

Considerable research efforts have been invested in improving the quality of the assembled graphs. The current state of the art techniques for these steps is now such that in many cases, simple (and efficient) algorithms and heuristics produce quite satisfying results on the assembled graphs.

It is important to keep in mind the main goal of broad-topic World Wide Web searches, which is to enhance the precision at 10 of the results, not to rank the entire collection of sites correctly. It is entirely irrelevant if the site in place 98 is really better than the site in place 216. The stochastic ranking, which turns out to be equivalent to a weighted in-degree ranking, discovers the most authoritative sites quite effectively (and very efficiently) in many (carefully assembled) collections. No claim is made on the quality of its ranking on the rest of the sites (which constitute the vast majority of the collection).

7. Conclusions

We have developed a new approach for finding hubs and authorities, which we call SALSA: the stochastic approach for link structure analysis. SALSA examines random walks on two different Markov chains which are derived from the link structure of the World Wide Web: the authority chain and the hub chain. The principal community of authorities (hubs) corresponds to the sites that are most frequently visited by the random walk defined by the authority (hub) Markov chain. SALSA and Kleinberg's mutual reinforcement approach are both in the framework of the same meta-algorithm.

We have shown that the ranking produced by SALSA is equivalent to a weighted in/out-degree ranking (with the sizes of irreducible components also playing a part). This makes SALSA computationally lighter than the mutual reinforcement approach.

Both approaches were tested on the World Wide Web, where SALSA appears to compare well with the mutual reinforcement approach. These tests, as well as analytical work, have revealed a topological phenomenon on the Web called the TKC effect. This effect sometimes derails the mutual reinforcement approach, and prevents it from finding relevant authoritative sites (or from finding authorities on all meanings/aspects of the query):

(1) In multi-topic collections, the principal community of authorities found by the mutual reinforcement approach tends to pertain to only one of the topics in the collection.

(2) In single-topic collections, the TKC effect sometimes results in the mutual reinforcement approach ranking many irrelevant sites as authorities.

We note that SALSA is less vulnerable to the TKC effect, and produces good results in many cases where the mutual reinforcement approach fails to do so.

The following issues are left for future research.

(1) In collections with many connected components, we have studied one manner in which to combine the inner-component authority score with the size of the component. There may be better ways to combine these two factors into a single score.

(2) We have found a simple property of the stochastic ranking, which enables us to compute this ranking without the need to approximate the principal eigenvector of the stochastic matrix which defines the random walk. Is there some simple property which will allow us to calculate the mutual reinforcement ranking without approximating the principal eigenvector of $W^{\mathrm{T}}W$? If not, can we alter the graph G in some simple manner (for instance, by changing some weights on the edges) so that the stochastic ranking on the modified graph will be approximately equal to the mutual reinforcement ranking on the original graph?

Acknowledgements

The second author would like to thank Udi Manber for introducing him to the search problems studied in this paper, and Udi Manber and Toni Pitassi for delightful and interesting discussions at the early stages of this research.

References

[1] J.G. Auguston and J. Minker, An analysis of some graph theoretical cluster techniques, JACM 17 (4) (1970) 571–588.

[2] S. Brin and L. Page, The anatomy of a large-scale hypertextual Web search engine, Proc. 7th Int. WWW Conf., 1998.

[3] J. Carrière and R. Kazman, Webquery: searching and visualizing the web through connectivity, Proc. 6th Int. WWW Conf., 1997.

[4] S. Chakrabarti, B. Dom, D. Gibson, J. Kleinberg, S.R. Kumar, P. Raghavan, S. Rajagopalan and A. Tomkins, Hypersearching the Web, Sci. Am. June 1999.

[5] S. Chakrabarti, B. Dom, D. Gibson, J. Kleinberg, S.R. Kumar, P. Raghavan, S. Rajagopalan and A. Tomkins, Mining the Web's link structure, IEEE Comp. August 1999.

[6] S. Chakrabarti, B. Dom, D. Gibson, S.R. Kumar, P. Raghavan, S. Rajagopalan and A. Tomkins, Spectral filtering for resource discovery, ACM SIGIR Workshop on Hypertext Information Retrieval on the Web, 1998.

[7] S. Chakrabarti, B. Dom, D. Gibson, J.M. Kleinberg, P. Raghavan and S. Rajagopalan, Automatic resource list compilation by analyzing hyperlink structure and associated text, Proc. 7th Int. WWW Conf., 1998.

[8] Compaq Computer Corporation, Altavista Net Guide, http://www.altavista.com/.

[9] R.G. Gallager, Discrete Stochastic Processes, Kluwer, Dordrecht, 1996.

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

[11] D. Gibson, J.M. Kleinberg and P. Raghavan, Inferring Web communities from link topology, Proc. 9th ACM Conf. on Hypertext and Hypermedia, 1998.

[12] Google Inc., Google Search Engine, http://www.google.com/.

[13] M.R. Henzinger and K. Bharat, Improved algorithms for topic distillation in a hyperlinked environment, Proc. 21st Int. ACM SIGIR Conf. on Research and Development in IR, August 1998.

[14] IBM Corporation Almaden Research Center, Clever, http://www.almaden.ibm.com/cs/k53/clever.html.

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

[16] J.M. Kleinberg, Authoritative sources in a hyperlinked environment, Proc. 9th ACM–SIAM Symp. on Discrete Algorithms, 1998.

[17] J.M. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan and A.S. Tomkins, The Web as a graph: measurements, models and methods, Proc. 5th Int. Computing and Combinatorics Conf., 1999.

[18] K. Law, T. Tong and A. Wong, Automatic Categorization Based on Link Structure, Stanford University, Stanford, 1999.

[19] M. Marchiori, The quest for correct information on the web: hyper search engines, Proc. 6th Int. WWW Conf., 1997.

[20] C.H. Papadimitriou, P. Raghavan, H. Tamaki and S. Vempala, Latent semantic indexing: a probabilistic analysis, Preliminary version appeared in PODS 98, pp. 159–168.

[21] P. Pirolli, J. Pitkow and R. Rao, Silk from a sow's ear: extracting usable structures from the Web, Proc. ACM SIGCHI Conf. on Human Factors in Computing, 1996.

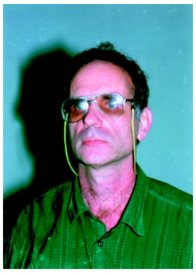
[22] H. Small, Co-citation in the scientific literature: a new measure of the relationship between two documents, Journal of the American Society for Information Science 24 (1973) 265–269.

[23] C.J. van Rijsbergen, Information Retrieval, Butterworths, 1979.

[24] R. Weiss, B. Vélez, M. Sheldon, C. Namprempre, P. Szilagyi, A. Duda and D. Gifford, Hypursuit: a hierarchical net-

work search engine that exploits content-link hypertext clustering, Proc. 7th ACM Conf. on Hypertext, 1996.

Ronny Lempel is a Ph.D. student in the Department of Computer Science, Technion, Haifa, Israel, focusing on World Wide Web link-structure analysis. He received his B.Sc. and M.Sc. from the same department in 1997 and 1999, respectively.



Shlomo Moran received his BSc and DSc degrees in mathematics from the Technion, in 1975 and 1979, respectively. Since 1981 he is a faculty member in the Computer Science Department in the Technion, where he is now a Professor and Chairman. His current research interests include communication in high speed networks, exact communication complexity of distributed tasks, confidentiality protection in medical records, and search methods in the Internet.