Introducing Regulated Bias into Co-Citation Ranking Schemes on the Web

Ronny Lempel
Department of Computer Science, The Technion, Haifa 32000, Israel.
Email: rlempel@cs.technion.ac.il

Shlomo Moran
Department of Computer Science, The Technion, Haifa 32000, Israel.
Email: moran@cs.technion.ac.il

Link structure analysis is widely recognized as an important tool in WWW information retrieval and is extensively used in the retrieval and ranking algorithms of major WWW search engines. Common extensions of link analyses involve changing the Web subgraphs that are analyzed. In this work, we present two techniques which bias co-citation based link analyses towards favorable pages and away from undesired pages. Both techniques are integrated seamlessly into existing link-analyzing algorithms.

Our first technique extends the notion of hubs by introducing *pseudo hubs*, which represent any trait of Web pages that is desirable to users. Pseudo hubs are added as pseudo nodes to the Web-graph and are linked to those pages which exhibit the desirable traits. The ensuing co-citation analysis then treats these pseudo nodes as if they were regular pages. This results in the rankings of pages with favorable traits being boosted. The second technique incorporates negative relevance judgments into the analysis, causing the results to shift away from the surroundings of undesired pages.

Introduction

Searching the WWW is an everyday task for millions of people around the world. People search for content of various media types and on every possible walk of life, relying on *search engines* to help them locate their points of interest from amongst the vast wealth of resources on the Internet (Lawrence and Giles, 1998).

Perhaps the most basic of search operations performed on the Web is querying for relevant textual content. Typically, the major search engines have used content-based analysis to serve such queries, matching the terms appearing in each query to the textual contents of Web pages. However, the performance of pure content-based search has significantly degraded as the Web grew larger and as *spamming* and *search engine persuasion* techniques became widely used (Marchiori, 1997; Chakrabarti et al., 1999a). The degradation of pure content-based searches has prompted research aimed at finding additional means for locating and ranking authoritative resources on the Web.

In recent years, the *link-structure* of the Web has been acknowledged as a rich source of information in search and retrieval problems on the Web (Marchiori, 1997). Kleinberg (Kleinberg, 1999) distinguished between two kinds of Web pages which pertain to a given query, which he named *hubs* and *authorities*. The authorities are those pages which contain the authoritative information pertaining to the query. Hubs, on the other hand, are primarily resource lists, linking to many authorities on the topic, possibly without directly containing the authoritative information. Kleinberg also devised an algorithm which finds pages that are good candidates for being hubs and authorities for a particular topic. That algorithm, which is based on co-citation of Web pages, has sparked IBM's *Clever* project (IBM) and has influenced many other works (such as (Bharat and Henzinger, 1998; Lempel and Moran, 2001)). Using a different approach, Brin and Page devised *PageRank* (Brin and Page, 1998), which produces a global importance measure of Web pages. The conclusion from all these works is that link-structure analysis (LSA) can vastly improve search and rank techniques which attempt to cope with topics of broad interest on the Web. Accordingly, search engines (such as *Google*[1] and *AltaVista*[2]) nowadays incorporate extensive LSA in their ranking algorithms, which determine the order in which the search results are displayed to the user (Brin and Page, 1998).

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

However, link structure analysis is not a silver bullet for solving all precision problems in Web searches. It has been noted that LSA sometimes suffers from *topic drift* (Chakrabarti et al., 1999b), may be fooled by *mutually reinforcing relationships between hosts* (Bharat and Henzinger, 1998) and may return irrelevant results due to densely interconnected sets of irrelevant pages (see the *TKC effect* in (Lempel and Moran, 2001)). Many research efforts were aimed at solving these problems, as well as at effectively combining LSA with additional relevance factors, particularly with the traditional content analysis of Web pages.

Of the proposed solutions, many have addressed the challenges above by computing several relevance scores for every page, and then combining the separate scores into a final score. The most common sub-scores found in the literature are text-based scores and link-based scores (e.g. (Aridor et al., 2000; Jin and Dumais, 2001; Marchiori, 1997)). The latter two works defined the link-based score of a page as a function of the text scores of its neighbors. The manner in which sub-scores are combined into a final ranking has in itself been the topic of recent research (Dwork et al., 2001).

A different approach to tackling the shortcomings of LSA is to analyze *modified* link structures. In such schemes, the connectivity patterns of the Web are manipulated in a manner which reflects additional information (like textual content) and which favorably biases the analysis. The modifications performed on link structures have mostly been of two types:

- Deleting irrelevant elements from the graph being analyzed. Links which are not seen as authority conferring, such as intra-domain links (Kleinberg, 1999), links to cgi-scripts (Brin and Page, 1998), advertisement links (Lempel and Moran, 2001) and others, are deleted by many algorithms. Nodes of pages that are ruled as irrelevant to some topic-specific search may also be deleted (Bharat and Henzinger, 1998), along with all their incident links.

- Assigning positive weights to the remaining links. For query specific searches, links with anchor text which is relevant to the query are often awarded high weights (Chakrabarti et al., 1998). Links which are incident to pre-defined pages of high quality may have their weights boosted (Chakrabarti et al., 1999b; Aridor et al., 2000). The influence of links may also be regulated according to the textual contents of their endpoints (Bharat and Henzinger, 1998), and links from many pages of the same host which point to many pages on another host may have their weights lowered (Bharat and Henzinger, 1998; Kleinberg et al., 1999).

To summarize, we see that pages and links which are deemed as irrelevant are normally ignored in the analysis, while textual information and other relevance factors influence the weights of the remaining links. These modifications to the actual link structure of the Web were shown in the mentioned works to produce improved results.

This work

We introduce two methods for integrating non-linkage based relevance factors into link structure analysis. These methods incorporate additional relevance information, both positive and negative, into the link structures which are the input to existing co-citation based ranking algorithms. In order to reflect the added information, we apply more changes to the actual topology of the Web. The changes, which include adding nodes (and links) and further manipulating the link weights, shift the produced rankings towards pages with favorable traits, and away from pages having negative qualities.

We first extend the notion of hubs by introducing *pseudo hubs*, which represent any trait of Web pages that is desirable to users. Pseudo hubs are added as pseudo nodes to the Web-graph and are linked to those pages which exhibit the desirable traits. The ensuing co-citation analysis treats the pseudo nodes as if they were regular pages, which results in the rankings of pages with favorable traits being boosted.

Next we consider negative relevance judgments, and how both static and dynamic knowledge of irrelevant pages can be integrated into LSA in a manner that will help us identify other irrelevant pages. We show that naively assigning negative weights[3] to links incident to irrelevant pages is unproductive. We proceed by presenting a reformulation of link weights which punishes pages that are related to the known irrelevant pages. We then demonstrate how our reformulation of link weights helps tune ambiguous search results to the information needs of different users.

Co-Citation Based Ranking Schemes

This section provides a brief overview of two co-citation based link analyses, Kleinberg's HITS algorithm (Kleinberg, 1999; Gibson et al., 1998), and SALSA (Lempel and Moran, 2001). Both techniques are *query specific, local* ranking algorithms, which first determine a small portion of the Web where resources on a specific query are likely to exist, and then analyze the link structure of that Web portion in order to assign query specific authority scores to pages in that portion alone.

Throughout this section, recall our assumption that every hyperlink $p \to q$ is assigned a weight $w(p \to q)$.

[3]Throughout this paper we will consider all links to have a *weight* associated with them.

Kleinberg's Hyperlink-Induced Topic Search (HITS)

HITS (Kleinberg, 1999; Gibson et al., 1998) starts by assembling a collection $\mathcal{C}$ of Web pages, which should contain communities of hubs and authorities pertaining to a given topic t. Denote by q a term-based search query which describes t. The collection $\mathcal{C}$ is assembled in the following manner:

- A *root set* S of pages is obtained by submitting the query q to a general WWW search engine.
- From S we derive a *base set* $\mathcal{C}$ which consists of (a) pages in the root set S, (b) pages which point to a page in S and (c) pages which are pointed to by a page in S.

Each page $s \in \mathcal{C}$ is now assigned a pair of weights, a hub-weight $h(s)$ and an authority weight $a(s)$, using the following iterative algorithm:

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

The top ranking pages, according to both kinds of weights, form the communities of hubs and authorities which HITS outputs.

The collection $\mathcal{C}$ and its (weighted) link structure induce a $|\mathcal{C}| \times |\mathcal{C}|$ adjacency matrix $W = [w_{i,j}]$, defined as follows:

$$w_{i,j} = \begin{cases} w(i \rightarrow j) & \text{page } i \text{ contains a link to page } j \\ 0 & \text{otherwise} \end{cases}$$

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 [4] of $W^T W$ [of WW^T].

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

SALSA

SALSA, the Stochastic Approach for Link Structure Analysis, is based on the stochastic properties of certain random walks that are performed on a given page collection.

The input to SALSA consists of a collection of pages $\mathcal{C}$ which is built around a topic t in the same manner as in HITS. From $\mathcal{C}$ and its link structure, SALSA builds the following bipartite undirected graph $G = (H, A, E)$ (Figure 1 exemplifies this construction):

- $H = \{s_h \mid s \in \mathcal{C} \text{ and } out-degree(s) > 0\}$ (the *hub-side* of G).
- $A = \{s_a \mid s \in \mathcal{C} \text{ and } in-degree(s) > 0\}$ (the *authority-side* of G).
- $E = \{(s_h, r_a) \mid s \rightarrow r \text{ in } \mathcal{C}\}$, with link weights being assigned to the corresponding edges.

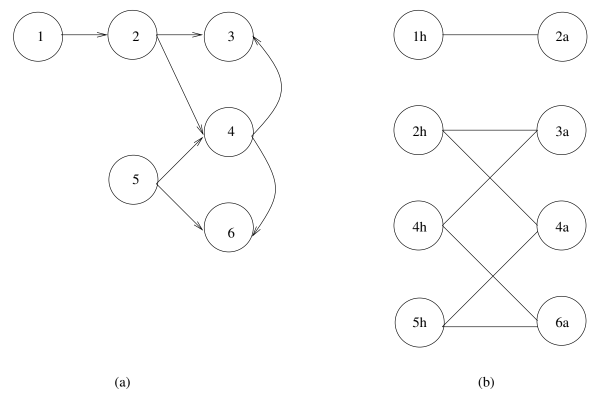

Figure 1: Transforming (a) the collection $\mathcal{C}$ into (b) a bipartite graph G

SALSA analyzes two different random walks on G. Each walk visits nodes from just one of the two sides of G, by traversing two G-edges in each step. In effect, state transitions in these random walks are equivalent to the traversal of two WWW-links in a row, one link forward and one link backwards (or vice versa). Analyzing both walks allows SALSA to give each Web page two distinct scores, a hub score and an authority score, which reflect the portion of time which the random walks spend in each node (page). Since the hubs and authorities of topic t should be highly visible in G, SALSA expects that the t-authorities will be amongst the nodes most frequently visited by the random walk on A, and that the t-hubs will be amongst the nodes most frequently visited by the random walk on H.

As mentioned, each state transition is composed of two edge crossings in G. Each single crossing is drawn independently according to the following rule: when leaving the G-node a, the probability of crossing the edge a-b equals $w(a-b)/E_a$, where E_a is the sum of weights of all the edges which are incident to a.

These random walks have an analytical solution (Lempel and Moran, 2001). Denote by $C(v)$, $v \in G$ the connected component of G which contains v. SALSA assigns each page $p \in \mathcal{C}$ hub and authority scores ($h(p)$ and $a(p)$ respectively) as follows:

$$
\begin{aligned}
a(p) &= \frac{E_{p_a}}{\sum_{v \in C(p_a) \cap A} E_v} \cdot \frac{|C(p_a) \cap A|}{|A|} \\
h(p) &= \frac{E_{p_h}}{\sum_{v \in C(p_h) \cap H} E_v} \cdot \frac{|C(p_h) \cap H|}{|H|}
\end{aligned}
$$

Pseudo Hubs

Pseudo hubs are means by which content-term information, user preferences and user relevance feedback can be integrated into link-analyzing algorithms. Pseudo hubs are entered as pseudo nodes into the link-topology graph of the WWW. They do not represent actual Web-pages, but rather certain characteristics which are deemed to be valuable for the information retrieval task at hand. Links are added from the pseudo nodes towards all of the Web pages which display the property which the pseudo node represents. These (pseudo) links may be weighted according to the presence of the tested property in each page. Thus, a pseudo node becomes a hub for its specific property (or characteristic).

For example, pseudo hubs can be used to emphasize keywords or terms by linking to pages which contain the query (or part of it), with link weights that reflect some query-page similarity measure. Other pseudo hubs may represent one or several domains (such as *.edu, .com, .gov*) from which users prefer to get results. They can also represent some property of the design of Web pages, and link to attractively designed pages (for example, users may prefer pages which contain images, animation or multimedia features over pages which contain primarily text even if the text-containing pages convey more raw information). A user's bookmarks file, or a topic-specific folder within that file, may constitute a pseudo-hub as well, reflecting the user's topic(s) of interest. Thus, pseudo hubs can incorporate multiple relevance factors into link-structure analysis.

In the analysis of both HITS and SALSA[5], pseudo hubs increase the authority weights of the pages to which they link, just as real hubs do. In SALSA, the relative ranking of other pages remains unaffected by pseudo hubs (the *rank local* property, defined in (Borodin et al., 2001)). In HITS, pseudo hubs will also boost the hub rankings of pages which co-cite pages to which the pseudo hubs points, and the effect will propagate further, by the $\mathcal{I}$ and $\mathcal{O}$ operations, to other pages as well. An important issue is how to assign weights to the links which emanate from the pseudo hubs. Choosing weights for these links is equivalent to *pricing* the property, or characteristic, which the pseudo hub represents ("how many links is the property worth"). Clearly, the weight assignments directly affect the bias which the pseudo hubs will introduce into the ranking algorithm. As the notions of relevance vary from user to user, so will the choice of weights assigned to the emanating links of the various pseudo hubs.

Another factor which influences the effect of a pseudo hub on its neighborhood is the number of non-pseudo links in the collection. If the collection contains many links, or at least the leading authorities are densely pointed at by real hubs, the effect of adding several pseudo links will be small. However, in sparse collections where the link analysis scores are based upon very few links, pseudo hubs and their outgoing links can change the rankings significantly. Effectively, the bias that pseudo-hubs introduce into the analysis grows when there are relatively few "real" links in the subgraph, and vice-versa. Intuitively, this seems as a positive feature of this approach, since LSA becomes more reliable and robust as the number of informative links in the collection grows. The more links are present, the higher is our confidence in the link-based scores and the less is our need for other indications of authoritativeness. However, when the collection is sparse, our confidence in the quality of link-based scores diminishes. We then turn to the relevance factors represented by the pseudo hubs, and allow them to take a more dominant part in the analysis.

The idea of tuning the dominance of link-based scores by the density of the collection was applied in (Aridor et al., 2000), where the score of each page was composed of text-based and link-based sub-scores. The two sub-scores were combined using an *adaptive linear combination*, in which the coefficient of the link-based score grew (roughly) in proportion to the amount of links in the collection. Again, the motivation was that LSA becomes less reliable when based upon relatively few links, and then text-based scores should dominate the rankings of the pages.

Pseudo Hub Examples

Our first example in Table 1 brings the top 5 authorities of a certain collection under two analyses with HITS. The first analysis did not include pseudo hubs, while the second included a pseudo hub which linked to the non-commercial pages of the collection (pages whose domains do not end with .com or .co.*).

The two result sets intersect in 3 of the 5 results. The first result in both cases, *www.searchenginewatch.com*, is indeed a search-engine technology site. Without pseudo hubs, all authorities ranked in places 2-20 are all commercial search engines or directories. In particular,

[5] The analysis holds whenever the addition of the pseudo links does not change the connected components of the graph G in SALSA, or equivalently, does not change the irreducible components of the matrix W^TW in HITS. In particular, the analysis holds whenever G was connected[W^TW was irreducible] before adding the pseudo links.

Table 1: Results for "Search Engine Technology", with and without pseudo hubs

```
Query: Search Engine Technology
Size of collection = 1353 pages

Pure: HITS analysis without pseudo hubs
PH:   HITS analysis with a domain-oriented
      pseudo hub, which links to all non-.com
      (or .co.*) pages
```

url (title)	Pure	PH
http://www.searchenginewatch.com (Search Engine Watch: Tips ...)	1	1
http://www.excite.com (Excite Start Page)	2	2
http://dmoz.org (ODP - Open Directory Project)		3
http://www.hotbot.com (HotBot)	3	
http://www.lycos.com (Lycos)	4	4
http://www.yahoo.com (Yahoo!)	5	
http://library.albany.edu/internet/search.html (Searching the Internet: Recommended ...)		5

the analysis was completely dominated by commercial pages. However, when adding a pseudo hub which favors non-commercial sites, the Open Directory project (*dmoz.org*) and an academic site on search techniques enter the top 5 authorities which HITS returned.

Our second example, shown in table 2, shows how a text-oriented pseudo hub can help in battling topic drift with HITS. On the query "high performance processors", the authorities found without pseudo hubs drifted to processor and chip manufacturers (the top 5 are shown, and the rest of the top 20 are of similar nature). We then added a pseudo hub that linked to pages with text that was highly similar to the query. The new top authorities shifted from general processor pages to pages on high performance processors and on processor comparison and evaluation.

The two top-20 lists do not intersect at all. The difference between the two result sets is perhaps best exemplified by the *Intel* page that belongs to each set. Without pseudo hubs, Intel's home page is ranked second among all pages. With the text-oriented pseudo hub, HITS found the OverDrive home page, which upgrades Intel's Pentium Pro processors.

LSA with Negative Relevance Judgments

In this section we examine how we can use negative relevance judgments on (actual or pseudo) pages in order to bias LSA away from the neighborhoods of those pages. The negative relevance judgments may be either *static*, query and user independent judgments, or *dynamic*, query and user specific judgments. Examples of the former, where search engines have a-priori knowledge on the undesirability of certain pages include the following:

- Search engines provide family filters, which exclude adult sites from the returned results. We may want to strengthen these filters by lowering the ranks of non-adult pages which point to adult sites.

- Search engines may want to bias their searches away from known spammers and their neighboring pages (which may also be spammers).

Table 2: Results for "High performance processors", with/without pseudo hubs

```
Query: High performance processors
Size of collection = 1128 pages

HITS analysis without pseudo hubs
```

url and title
http://www.amd.com/ AMD - Advanced Micro Devices...
http://www.intel.com/ Welcome to Intel
http://www.cyrix.com/ VIA Technologies, Inc.
http://www.ti.com/ Texas Instruments Welcomes You
http://www.analog.com/ Analog Devices, Inc.

```
HITS analysis with a text-oriented pseudo hub,
which points to pages with text that is highly
similar to the query.
```

url, title
http://www.pcmech.com/cpuindex.htm Processors
http://carol.wins.uva.nl/~andy/apr/node4.html High performance processors
http://www.intelligentfirm.com Processor Performance, Evaluation and Programming
http://www.sysopt.com/cpu.html Processor Information
http://www.digit-life.com/articles/iwillka266randviaes DDR chipsets investigated: ALi MAGiK 1 and VIA Pro266
http://www.intel.com/overdrive OverDrive Processor Home
http://company.monster.com/alpro ALPHA-PROCESSOR

Dynamic relevance feedback is obtained by the engines by tracking the actions of users when search results are returned to them. Such tracking mechanisms can be implemented by a combination of cookies, javascript and cgi scripts, as demonstrated by the *SearchPad* system (Bharat, 2000). The engines may obtain dynamic negative relevance feedback either explicitly or implicitly:

- Engines may explicitly collect negative relevance feedback by allowing users to voice their objection to the results returned by the engine. For example, the engines may complement the already available "more like this"/"similar pages" option with a "not like this" option next to every returned result.

- Implicit negative feedback may be inferred by the engines by tracking which returned results were *not viewed* by a user, under the assumption that if a result (URL and short description) did not prompt the user to follow the link to the returned URL, then that result is deemed irrelevant by the user[6].

Both explicit and implicit negative feedback can be taken into account when the user requests additional result pages for the same query. We will demonstrate such scenarios later in this Section.

We now formally describe our treatment of negative feedback. Let $\mathcal{C}$ be a collection which is built around a topic t, and let $\mathcal{B} \subset \mathcal{C}$ be a subset of *bad* pages which are known to be irrelevant to t. Let $\mathcal{G} \stackrel{\triangle}{=} \mathcal{C} \setminus \mathcal{B}$ denote the set of potentially *good* pages, which we want to rank.

As discussed in the Introduction, the standard practice today is to extract $\mathcal{B}$-pages (and their incident links) from the graph which is analyzed. While extracting (in- and out-) links that are incident to irrelevant pages does indeed lower the rankings of the neighbors of those pages, it seems that those neighbors should suffer a more severe penalty to their scores, as the example in figure 2 indicates: Assume that page B is known to be

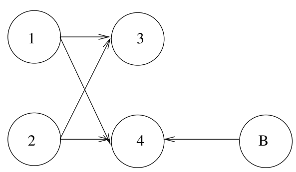

Figure 2: Links from irrelevant pages affect our perception of authority

irrelevant. By simply erasing the link $B \to 4$, pages 3 and 4 become indistinguishable as possible authorities in the collection. However, intuition suggests that the presence of the link $B \to 4$ should cause page 3 to be ranked as a better authority than page 4. We are thus motivated to explore how knowing the identities of the $\mathcal{B}$-pages can help to bias the analysis away from the neighborhoods of those pages, and to better rank the $\mathcal{G}$-pages. Specifically, we will try to find a weighting of edges which will hurt the rankings of pages in $\mathcal{G}$ with neighbors in $\mathcal{B}$.

Negative Link Weights

We have mentioned the assignment of large positive weights to links as a technique which improves the rankings of the pages which are incident to those links. A natural approach which follows is the assignment of negative weights to links connecting potentially good pages and known bad ones. In this subsection we discuss this approach in the context of the co-citation algorithms HITS and SALSA, and conclude that even though the idea is intuitive, the approach is unproductive. We first observe that SALSA cannot be formulated in terms of negative link weights, since these result in undefined, negative transition probabilities. In what follows, we mention three natural ways of inserting negative link weights into HITS, and discuss how two of these will affect HITS.

Given a collection whose pages are partitioned into potentially good ($\mathcal{G}$) pages and known bad ($\mathcal{B}$) pages, and whose links all have positive weights, we set forth by negating the weights of links which connect $\mathcal{G}$-pages with $\mathcal{B}$-pages, while not changing the weights of links which connect two $\mathcal{G}$-pages. As for links which connect two $\mathcal{B}$-pages, three reasonable approaches naturally emerge: (1) leaving their weights as is, (2) negating their weights, or (3) erasing those links from the collection. Formally, denoting by $W = [w_{i,j}]$ the (non-negative) adjacency matrix of the collection, our three approaches correspond respectively to the matrices W_1, W_{-1} and W_0 defined below:

$$W_k = [w_{k\ i,j}] = \begin{cases} -w_{i,j} & i \in \mathcal{B}, j \in \mathcal{G} \text{ or } i \in \mathcal{G}, j \in \mathcal{B} \\ w_{i,j} & i,j \in \mathcal{G} \\ k \cdot w_{i,j} & i,j \in \mathcal{B} \end{cases}$$

Note that we do not actually care about the rankings of the bad pages themselves, since we can always ignore them (we know their identities). The importance of the weights assigned to links between bad pages lies only in their indirect effect on the rankings of good pages.

Next we show that while HITS can be applied using each of the three possibilities of new edge weights, this is not helpful in the cases of W_1 and W_0. Specifically, we claim that:

1. Applying HITS with W_1 will assign the $\mathcal{G}$-pages the same weights as will applying it with (the original) W .
2. Applying HITS with W_0 will assign the $\mathcal{G}$-pages the same weights as will (1) pre-processing the collection by deleting all links which have both endpoints in $\mathcal{B}$, and (2) applying HITS on (the deletion-induced adjacency matrix) W.

The second claim is an immediate consequence of the first. We prove the first result for the authority weights. The proof for the hub weights is similar.

Proposition 1 *Let $M = [m_{i,j}]$ be the following $|\mathcal{C}| \times |\mathcal{C}|$ matrix:*

$$m_{i,j} = \begin{cases} 1 & i = j \in \mathcal{G} \\ -1 & i = j \in \mathcal{B} \\ 0 & i \neq j \end{cases}$$

Then, v is an eigenvector of $W^T W$ which corresponds to the eigenvalue λ if and only if Mv is an eigenvector of $W_1^T W_1$ which corresponds to λ.

proof: First, observe that $M = M^{-1} = M^T$: M is diagonal, and all of its entries on the main diagonal have an absolute value of 1. From the definition of the

[6] Note, however, that users occasionally do not follow links even though they identify the links as relevant, such as links they are already familiar with. Similarly, not all followed links lead to relevant pages, as disappointed users sometimes find out.

transformations of the weights, it is easy to see that $W = MW_1M$, and hence $W^TW = MW_1^TMMW_1M = MW_1^TW_1M$. Thus, the matrices $W_1^TW_1$ and W^TW are similar, and the proposition follows (Lancaster and Tismenetsky, 1985). $\square$

Since the authority weights in HITS correspond to eigenvectors of W^TW (or of $W_k^TW_k$), it follows that the only authority weights which change by using W_1 instead of W are those which correspond to the pages of $\mathcal{B}$, which are known to be bad in the first place. No change is observed in the weights of any $\mathcal{G}$-page, whether in the neighborhood of the $\mathcal{B}$-pages or not. As noted in page 6, erasing the links incident to bad pages lowers the rankings of the neighbors of those pages. Consequently, when applying HITS, erasing links that are incident to bad pages is a better strategy than using either of the weight transformations W_1 and W_0. The qualitative behavior of HITS using W_{-1} appears to be less clear, and needs to be further explored.

Modifying Link Weights In Light of Negative Judgments

We are interested in finding link weights which will prefer $\mathcal{G}$-pages with (potentially) respectable neighbors over $\mathcal{G}$-pages with many neighboring bad pages. However, as mentioned in the previous subsection, naively negating the weights of edges connecting good and bad pages is unproductive. In this subsection we pursue an approach which deletes all bad pages (and their incident links), and reduces the (positive) weights of the incident links of $\mathcal{G}$-pages which had bad neighbors. We first define our approach mathematically, and then present several scenarios that demonstrate its potential.

We proceed to define the in- and out- *fidelities* of a page $p \in \mathcal{G}$, which measure the proportion of the weights of p's (incoming/outgoing) links to other $\mathcal{G}$-pages out of the weights of all of p's incident links:

$$f_{in}(p) = \frac{d_{in}^+(p)}{d_{in}^+(p) + d_{in}^-(p)} \quad , \quad f_{out}(p) = \frac{d_{out}^+(p)}{d_{out}^+(p) + d_{out}^-(p)}$$

where $d^+(\cdot)$ and $d^-(\cdot)$ are the following quantities:

$$\begin{aligned}
d_{in}^+(p) &\triangleq \sum_{\{k \in \mathcal{G} | k \to p\}} w(k \to p) \\
d_{in}^-(p) &\triangleq \sum_{\{k \in \mathcal{B} | k \to p\}} w(k \to p) \\
d_{out}^+(p) &\triangleq \sum_{\{k \in \mathcal{G} | p \to k\}} w(p \to k) \\
d_{out}^-(p) &\triangleq \sum_{\{k \in \mathcal{B} | p \to k\}} w(p \to k)
\end{aligned}$$

We now redefine the weights of the collection's links as follows:

$$w^*(h \to a) \triangleq \begin{cases} f_{out}(h) \cdot w(h \to a) \cdot f_{in}(a) & h, a \in \mathcal{G} \\ 0 & \text{otherwise} \end{cases}$$

The weights of links between $\mathcal{G}$-pages are altered according to the fidelities of both endpoints of the link. The weights of all other links is set to zero.

SALSA can now proceed normally with with the new, non-negative weights $w^*(\cdot,\cdot)$. These weights restrict SALSA's random walks to the set of $\mathcal{G}$-pages, since $\mathcal{B}$-pages have zero in/out degrees and are not represented in SALSA's bipartite graph. Substituting the weights $w^*(\cdot,\cdot)$ in the formulas for SALSA's scores, we conclude that SALSA will assign the following authority weights to each $p \in \mathcal{G}$[7]:

$$\sum_{\{h \in \mathcal{G} | h \to p\}} w^*(h \to p) \;=\; f_{in}(p) \sum_{\{h \in \mathcal{G} | h \to p\}} f_{out}(h) \cdot w(h \to p)$$

The SALSA hub weight of each $p \in \mathcal{G}$ will be

$$\sum_{\{a \in \mathcal{G} | p \to a\}} w^*(p \to a) \;=\; f_{out}(p) \sum_{\{a \in \mathcal{G} | p \to a\}} w(p \to a) \cdot f_{in}(a)$$

These modified scores have an appealing reasoning: When summing over the incoming[outgoing] links of p, the influence of each link is scaled according to the fidelity of its other endpoint, lowering the impact on p of its $\mathcal{G}$-neighbors which have bad neighbors. The resulting sum is further scaled by p's own fidelity, punishing p for *its own* bad neighbors.

An alternative, more aggressive definition of the fidelities is the following:

$$f_{in}(p) \triangleq d_{in}^+(p) - d_{in}^-(p) \quad , \quad f_{out}(p) \triangleq d_{out}^+(p) - d_{out}^-(p)$$

These definitions may result in some fidelities being negative. We proceed by moving $\mathcal{G}$-pages with negative fidelities to $\mathcal{B}$, redefining $d^+(\cdot)$ and $d^-(\cdot)$ accordingly, and recalculating the fidelities. We may require several iterations of this process, but once arriving at a state where all remaining $\mathcal{G}$-pages have non-negative fidelities, the previous definition of the new link weights $w^*(\cdot,\cdot)$ is applied, and the analysis follows in the exact same manner.

The analysis of the behavior of HITS with the weights $w^*(\cdot,\cdot)$ is not included in this paper, although HITS is used in several of the example scenarios that follow.

Examples of Negative Relevance Feedback

We bring several scenarios that demonstrate how dynamic negative relevance feedback can be utilized. Each scenario starts by showing the top 10 results for an ambiguous or polarized query. We then consider two different (imaginary) users, who are interested in different

[7]The shown weights are further scaled by a constant factor within each connected component of SALSA's bipartite graph

aspects of the query. The two users thus view different subsets of the top 10 results. The results viewed by each user imply a negative opinion on the *unviewed* results, and this negative feedback is considered when ranking the next batch of results (results $11-20$) that each user will see. All the following examples are calculated using the first, less aggressive definition of the fidelities.

Table 3 shows a scenario for the query "jaguar", where user (a) is interested in the vehicle while user (b) is interested in video games. Both users view 3 relevant pages of the top 10 results. Then, in results $11-20$, user (a) receives 6 car pages and only a single video-game page, while user (b) receives 3 video-game pages and 5 car pages (and 3 of those 5 car pages are ranked in places $18-20$).

Table 4 shows a scenario for the query "football". User (a) is interested in professional American football and views just one of the top 10 results, while user (b) is interested in soccer and views 3 results. Then, in results $11-20$, the top 4 results user (a) receives are American football pages, while 4 of the top 7 results user (b) sees are soccer pages.

Table 5 shows a scenario for the query "gun control", where user (a) is searching for pro-gun control pages, while user (b) is searching for pages that oppose gun control. Both users view 3 of the top 10 pages, and they then receive disjoint sets for the results in places $11-20$. The new results returned to each user are focussed on that user's view of the topic, as well as on the legal aspects of gun control.

Table 6 brings a scenario for the query "abortion", where user (a) is searching for pro-choice pages while user (b) is searching for pro-life pages. 9 of the top 10 results are pro-choice, and so user (b) does not select any result, and receives 8 pro-life results in the next batch of 10 results. User (a) selects 3 of the pro-choice pages, and receives 8 additional pro-choice resources in the next batch of results.

It is possible to complement our treatment of unfollowed links as irrelevant pages with the treatment of followed links as a form of positive relevance feedback. Several works have integrated positive feedback in the form of known good resources on a topic into LSA. In (Chakrabarti et al., 1999b) and (Aridor et al., 2000), the weights of links that are incident to the known high-quality pages are boosted. This improves the rankings of pages in the vicinity of the designated pages. A more complex approach is used in (Chang et al., 2000), where every entry of the adjacency matrix W (the weight of every link in the collection) undergoes iterative adjustments. These adjustments align the principal community of authorities in the direction of the predefined page (or pages) of high quality. We chose not to apply any of these techniques in our example scenarios, since we wanted the scenarios to reflect the potential of the negative feedback approach alone. Finally, we note that there may be other methods of tuning the second

Table 3: The effect of negative links for the query "jaguar"

```
Query: ''jaguar''
LSA Algorithm: SALSA
```

The top 10 authorities:

url,title	topic, selected by
www.jag-lovers.org/ Jag-lovers - the Jaguar Enthusiasts'...	cars (a)
www.jaguarcars.com/ Jaguar Cars Global Home Page	cars (a)
www.jagweb.com/ A1 JagWeb - Jaguar restoration,...	cars
www.xke-lovers.com/ XK-E-Lovers	cars (a)
www.telegames.com/ Telegames Companies Home Page	video (b)
jaguar.holyoak.com/ Jagu-Dome, Jagu-Sounds and More!	video (b)
www.jec.org.uk/ Jaguar Enthusiasts' Club	cars
www.mossmotors.com/ Moss Motors	cars
www.magicnet.net/~jaguar/jcof.html Jaguar Club of Florida	Cars
www.vglq.com/ Video Game Liquidators...	video (b)

Authorities ranked $11-20$, user (a):

url,title	topic
www.terrysjag.com/ Terry's Jaguar Parts	cars
www.jags.org/ Jaguar Associates Group	cars
www.jaguarvehicles.com/ Jaguar Cars Global Home Page	cars
www.classicar.com/museums/welshjag/welshjag.html (URL no longer available)	cars
www.intex.net/~cci/index.html (URL no longer available)	other
users.aol.com/mikecoxtx/trombones/jagbones.htm LBJ Trombones	other
www.linkexchange.com/ Microsoft bCentral	other
www.hdrogers.com/ Jaguar parts auto parts...	cars
www.jag-lovers.org/xjlovers/ XJ Lovers...	cars
www.battlesphere.com BattleSphere Official Homepage	video

Authorities ranked $11-20$, user (b):

url,title	topic
www.terrysjag.com/ Terry's Jaguar Parts	cars
www.myatari.com/ Atari computers...	video
www.classicar.com/museums/welshjag/welshjag.htm (URL no longer available)	cars
www.battlesphere.com BattleSphere Official Homepage	video
www.linkexchange.com/ Microsoft bCentral	other
users.aol.com/mikecoxtx/trombones/jagbones.htm LBJ Trombones	other
www.atarihq.com/ Atari Gaming Headquarters...	video
www.jaguarvehicles.com/ Jaguar Cars Global Home Page	cars
www.jags.org/ Jaguar Associates Group	cars
www.hdrogers.com/ Jaguar parts auto parts...	cars

Table 4: The effect of negative links for the query "football"

Query: ''football''
LSA Algorithm: SALSA

The top 10 authorities:

url,title	topic, selected by
www.nfl.com NFL.com	A. Football (a)
www.fifa.com Federation Internationale de Football...	Soccer (b)
www.soccernet.com ESPN.com Soccernet: Home	soccer (b)
www.cfl.ca The Official Site of the Canadian...	Canadian Football
www.goonersguide.com The Gooners Guide...	soccer bets
www.arenafootball.com Welcome to Arena Football League	A. Football
www.ncaafootball.net ncaafootball – Welcome	A. Football
www.sportingnews.com SportingNews.com	American sports
www.football365.com Football news, fun, results...	soccer (b)
www.hqfl.dk High Quality Football Logos	soccer

Authorities ranked $11-20$, user (a):

url,title	topic
www.usatoday.com/sports/nfl.htm USATODAY.com	American football
football.espn.go.com/nfl/index ESPN.com:NFL	American football
www.profootballhof.com Welcome to the Pro Football Hall...	American football
www.sportingnews.com/nfl SportingNews.com : NFL	American football
www.the-fa.org The Football Association...	soccer
www.footballnews.co.uk FootballNews: Number 1 for...	soccer
www.TheCounter.com TheCounter.com:...	other
www.usafootball.com (untitled)	American football
www.football-punter.co.uk Football-Punter.co.uk	soccer bets
www.soccer-links.com soccer-links.com: The soccer...	soccer

Authorities ranked $11-20$, user (b):

url,title	topic
www.the-fa.org The Football Association...	soccer
www.footballnews.co.uk FootballNews: Number 1 for...	soccer
www.usatoday.com/sports/nfl.htm USATODAY.com	American football
www.planetfootball.com Main News - ...	soccer
www.TheCounter.com TheCounter.com:...	other
football.espn.go.com/nfl/index ESPN.com:NFL	American football
www.soccer-links.com soccer-links.com: The soccer...	soccer
www.sportingnews.com/nfl SportingNews.com : NFL	American football
www.profootballhof.com Welcome to the Pro Football Hall...	American football
www.usafootball.com (untitled)	American football

Table 5: The effect of negative links for the query "gun control"

Query: ''gun control''
LSA Algorithm: HITS

The top 10 authorities:

url,title	view, selected by
www.nra.org National Rifle Association - MyNRA	con (b)
www.handguncontrol.org The Brady Campaign to Prevent Gun...	pro (a)
www.jpfo.org Jews for the Preservation of Firearms...	con
www.vpc.org The Violence Policy Center	pro (a)
www.wagc.com Women Against Gun Control	con
www.guncite.com GunCite: gun control and Second...	con (b)
www.guntruths.com GunTruths: The truth about guns	con (b)
www.gunfree.org Coalition to Stop Gun Violence	pro (a)
www.citizensofamerica.org Citizens Of America - Pro-Gun Rights...	con
www.2ndlawlib.org Second Amendment Law Library	legal

Authorities ranked $11-20$, user (a):

url,title	view
www.jointogether.org Join Together Online...	pro
www.atf.treas.gov ATF Online - Bureau of...	legal
www.millionmommarch.com The Million Mom March...	pro
www.lcav.org LCAV Legal Community Against Violence	pro
www.americansforgunsafety.com Americans for Gun Safety...	pro
www.potomac-inc.org Second Amendment, Gun Control...	pro
support.jhsph.edu/departments/gunpolicy John Hopkins Center for Gun Policy and Research	legal
www.helpnetwork.org Welcome to The HELP Network	pro
www.ichv.org Illinois Council Against Handgun Violence...	pro
www.hgabuse.org About Marylanders Against Handgun Abuse	pro

Authorities ranked $11-20$, user (b):

url,title	view
www.sas-aim.org Second Amendment Sisters', Inc.	con
www.keepandbeararms.com Keep and Bear Arms - Gun Owners...	con
thomas.loc.gov THOMAS – U.S. Congress on the Internet	legal
www.house.gov United States House of Representatives...	legal
www.gunscholar.org The Journalist's Guide to Gun Policy	legal
www.senate.gov U.S. Senate...	legal
www.vote-smart.org Project Vote Smart	(politics)
www.calnra.org CalNRA.org	con
www.nraila.org (NRAILA)	con
www.packing.org Packing.org Concealed Carry: Home	con

Table 6: The effect of negative links for the query "abortion"

```
Query: ''abortion''
LSA Algorithm: HITS
```

The top 10 authorities:

url,title	view, selected by
www.naral.org NARAL: Abortion and Reproductive...	choice (a)
www.prochoice.org NAF - The Voice of Abortion Providers	choice (a)
www.gynpages.com Abortion Clinics OnLine	choice
www.plannedparenthood.org Planned Parenthood Federation...	choice (a)
www.cais.com/agm/main Ardent Communications	(other)
www.ms4c.org Medical Students for Choice	choice
www.caral.org California Abortion and Reproductive...	choice
www.feminist.org FEMINIST MAJORITY FOUNDATION...	choice
www.agi-usa.org The Alan Guttmacher Institute: Home...	neutral
www.ncap.com National Coalition of Abortion Providers	choice

Authorities ranked $11 - 20$, user (a):

url,title	view
www.nrlc.org National Right to Life	life
www.rcrc.org The Religious Coalition for Reproductive Choice	choice
www.nnaf.org National Network of Abortion Funds	choice
www.cath4choice.org Catholics for a Free Choice	choice
www.wcla.org Welcome to the Westchester Coalition for Legal...	choice
www.prochoiceresource.org ProChoice Resource Center	choice
www.aclu.org ACLU: American Civil Liberties Union	choice
ethics.acusd.edu/abortion.html Abortion; Ethics	neutral
www.ippf.org International Planned Parenthood Federation	choice
www.now.org/issues/abortion NOW and Abortion Rights...	choice

Authorities ranked $11 - 20$, user (b):

url,title	view
www.nrlc.org National Right to Life	life
www.afterabortion.org After Abortion: Information...	life
www.pregnancycenters.org Pregnancy Centers Online - Help...	neutral
www.prolifeinfo.org The Ultimate Pro-Life Resource List	life
www.hli.org (untitled)	life
www.abortionfacts.com Abortion facts and information...	life
www.lifecall.org Abortion Absolutely Not...	life
www.abortioncancer.com Abortion - Breast Cancer Link...	life
www.prolifeaction.org Pro-Life Action League	life
www.angelfire.com/tx2/abortion (not found)	(unknown)

result page of an ambiguous search to the preferences the user has shown while viewing the first result page. Specifically, the sets of followed/unfollowed first-page results may be coupled with a clustering of the ambiguous search results to produce a second result page that contains results from the cluster of the followed results and does not contain results from the clusters of the unfollowed ones.

Conclusions

The current practice for introducing bias into link structure analysis involves changing several aspects of the Web subgraphs that are analyzed, such as removing links and pages from the subgraphs, and assigning different positive weights to links as means of regulating the influence which these links have on the analysis.

We have studied techniques for further integrating general relevance factors into link structure analysis. These factors, which may be of either positive or negative nature, are integrated seamlessly into existing cocitation based link analyses. We are able to bias the analyses towards pages with favorable traits and away from undesired pages and their neighborhoods.

Our first novel technique introduces pseudo hubs as means of including general positive relevance factors into link analysis. Pseudo hubs are non-page entities which confer authority onto Web pages. Pseudo hubs are entered as pseudo nodes into the Web subgraph whose link structure is analyzed, as if they were regular resource lists. The information which they convey is thus integrated into the analysis. Information which can be represented by pseudo hubs includes lexical information and relevance judgments from any source. We have shown two examples of the influence of pseudo hubs in standard link-analyzing schemes. We have argued that the influence of pseudo-hubs on the analysis varies inversely to the amount of "real" links in the subgraph, which intuitively seems to be a positive feature of this approach.

Our second contribution examines how knowledge of some irrelevant pages can help LSA identify other irrelevant pages. We have shown that some natural schemes of assigning negative weights to the incident links of the known irrelevant pages are unproductive. We then suggested a reformulation of link weights which causes the ensuing analysis to shift away from undesired pages. We demonstrated several scenarios where this approach was used to dynamically tune ambiguous search results to the information needs of the users.

References

Aridor, Y., Carmel, D., Lempel, R., Maarek, Y. S., and Soffer, A. (2000). Knowledge agents on the web. *Proc. 4th International Workshop on Cooperative Information Agents.*

Bharat, K. (2000). Searchpad: explicit capture of search context to support web search. In *Proceedings 9th International World Wide Web Conference, Amsterdam*.

Bharat, K. and Henzinger, M. R. (1998). Improved algorithms for topic distillation in a hyperlinked environment. *Proc. 21'th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Borodin, A., Roberts, G. O., Rosenthal, J. S., and Tsaparas, P. (2001). Finding authorities and hubs from link structures on the world wide web. In *Proc. 10th International World Wide Web Conference*, pages 415–429.

Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Proc. 7th International WWW Conference*.

Chakrabarti, S., Dom, B., Gibson, D., Kleinberg, J., Kumar, S., Raghavan, P., Rajagopalan, S., and Tomkins, A. (1999a). Hypersearching the web. *Scientific American*.

Chakrabarti, S., Dom, B., Gibson, D., Kleinberg, J., Kumar, S., Raghavan, P., Rajagopalan, S., and Tomkins, A. (1999b). Mining the link structure of the www. *IEEE Computer*.

Chakrabarti, S., Dom, B., Gibson, D., Kleinberg, J. M., Raghavan, P., and Rajagopalan, S. (1998). Automatic resource list compilation by analyzing hyperlink structure and associated text. *Proc. 7th International WWW Conference*.

Chang, H., Cohn, D., and McCallum, A. (2000). Creating customized authority lists. In *Proc. 17th International Conference on Machine Learning (ICML-2000)*, pages 127–134.

Dwork, C., Kumar, R., Naor, M., and Savikumar, D. (2001). Rank aggregation methods for the web. In *Proc. 10th International World Wide Web Conference*, pages 613–622.

Gibson, D., Kleinberg, J. M., and Raghavan, P. (1998). Inferring web communities from link topology. *Proc. 9th ACM Conference on Hypertext and Hypermedia*.

IBM. IBM Corporation, Almaden Research Center. CLEVER. http://www.almaden.ibm.com/cs/k53/clever.html.

Jin, R. and Dumais, S. (2001). Probabilistic combination of content and links. In *Proc. 24'th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 402–403.

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *J. ACM*, 46:5:604–632.

Kleinberg, J. M., Kumar, R., Raghavan, P., Rajagopalan, S., and Tomkins, A. S. (1999). The web as a graph: Measurements, models and methods. *Proceedings of the Fifth International Computing and Combinatorics Conference*.

Lancaster, P. and Tismenetsky, M. (1985). *The Theory of Matrices*. Academic Press.

Lawrence, S. and Giles, C. L. (1998). Searching the world wide web. *Science*, 280.

Lempel, R. and Moran, S. (2001). SALSA: The stochastic approach for link-structure analysis. *ACM Transactions on Information Systems*, 19(2):131–160.

Marchiori, M. (1997). The quest for correct information on the web: Hyper search engines. *Proc. 6th International WWW Conference*.