Topic-Sensitive PageRank

Taher H. Haveliwala*
Stanford University
Computer Science Department
Stanford, CA 94305
taherh@cs.stanford.edu
(650) 723-9273

ABSTRACT

In the original PageRank algorithm for improving the ranking of search-query results, a single PageRank vector is computed, using the link structure of the Web, to capture the relative "importance" of Web pages, independent of any particular search query. To yield more accurate search results, we propose computing a *set* of PageRank vectors, biased using a set of representative topics, to capture more accurately the notion of importance with respect to a particular topic. By using these (precomputed) biased PageRank vectors to generate query-specific importance scores for pages at query time, we show that we can generate more accurate rankings than with a single, generic PageRank vector. For ordinary keyword search queries, we compute the topic-sensitive PageRank scores for pages satisfying the query using the topic of the query keywords. For searches done in context (e.g., when the search query is performed by highlighting words in a Web page), we compute the topic-sensitive PageRank scores using the topic of the context in which the query appeared.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*search process, information filtering, retrieval models*; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*linguistic processing*

General Terms

Algorithms, Experimentation

Keywords

search, Web graph, link structure, PageRank, search in context, personalized search

1. INTRODUCTION

Various link-based ranking strategies have been developed recently for improving Web-search query results. The HITS algorithm proposed in [14] relies on query-time processing to deduce the *hubs* and *authorities* that exist in a subgraph of the Web consisting of both the results to a query and the local neighborhood of these results. [4] augments the HITS algorithm with content analysis to improve precision for the task of retrieving documents related to a query topic (as opposed to retrieving documents that exactly satisfy the user's information need). [8] makes use of HITS for automatically compiling resource lists for general topics.

The PageRank algorithm discussed in [7, 16] precomputes a rank vector that provides a-priori "importance" estimates for all of the pages on the Web. This vector is computed once, offline, and is independent of the search query. At query time, these importance scores are used in conjunction with query-specific IR scores to rank the query results. PageRank has a clear efficiency advantage over the HITS algorithm, as the query-time cost of incorporating the *precomputed* PageRank importance score for a page is low. Furthermore, as PageRank is generated using the entire Web graph, rather than a small subset, it is less susceptible to localized link spam.

In this paper, we propose an approach that (as with HITS) allows the query to influence the link-based score, yet (as with PageRank) requires minimal query-time processing. In our model, we compute offline a *set* of PageRank vectors, each biased with a different topic, to create for each page a *set* of importance scores with respect to particular topics. The idea of biasing the PageRank computation was suggested in [6] for the purpose of *personalization*, but was never fully explored. This biasing process involves introducing artificial links into the Web graph during the offline rank computation, and is described further in Section 2.

By making PageRank topic-sensitive, we avoid the problem of heavily linked pages getting highly ranked for queries for which they have no particular authority [3]. Pages considered important in some subject domains may not be considered important in others, regardless of what keywords may appear either in the page or in anchor text referring to the page [5]. An approach termed *Hilltop*, with motivations similar to ours, is suggested in [5] that is designed to improve results for *popular* queries. Hilltop generates a query-specific authority score by detecting and indexing pages that appear to be good experts for certain keywords, based on their outlinks. However, query terms for which experts were not found will not be handled by the Hilltop algorithm.

[17] proposes using the set of Web pages that contain some

*Supported by NSF Grant IIS-0085896 and an NSF Graduate Research Fellowship.

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

term as a bias set for influencing the PageRank computation, with the goal of returning terms for which a *given* page has a high reputation. An approach for enhancing search rankings by generating a PageRank vector for each possible query term was recently proposed in [18] with favorable results. However, the approach requires considerable processing time and storage, and is not easily extended to make use of user and query *context*. Our approach to biasing the PageRank computation is novel in its use of a small number of representative basis topics, taken from the Open Directory, in conjunction with a unigram language model used to classify the query and query context.

In our work we consider two scenarios. In the first, we assume a user with a specific information need issues a query to our search engine in the conventional way, by entering a query into a search box. In this scenario, we determine the topics most closely associated with the query, and use the appropriate topic-sensitive PageRank vectors for ranking the documents satisfying the query. This ensures that the "importance" scores reflect a preference for the link structure of pages that have some bearing on the query. As with ordinary PageRank, the topic-sensitive PageRank score can be used as part of a scoring function that takes into account other IR-based scores. In the second scenario, we assume the user is viewing a document (for instance, browsing the Web or reading email), and selects a term from the document for which he would like more information. This notion of *search in context* is discussed in [10]. For instance, if a query for "architecture" is performed by highlighting a term in a document discussing famous building architects, we would like the result to be different than if the query "architecture" is performed by highlighting a term in a document on CPU design. By selecting the appropriate topic-sensitive PageRank vectors based on the context of the query, we hope to provide more accurate search results. Note that even when a query is issued in the conventional way, without highlighting a term, the *history* of queries issued constitutes a form of query context. Yet another source of context comes from the *user* who submitted the query. For instance, the user's bookmarks and browsing history could be used in selecting the appropriate topic-sensitive rank vectors. These various sources of search context are discussed in Section 5.

A summary of our approach follows. During the offline processing of the Web crawl, we generate 16 topic-sensitive PageRank vectors, each biased (as described in Section 2) using URLs from a top-level category from the Open Directory Project (ODP) [2]. At query time, we calculate the similarity of the query (and if available, the query or user context) to each of these topics. Then instead of using a single global ranking vector, we take the linear combination of the topic-sensitive vectors, weighted using the similarities of the query (and any available context) to the topics. By using a *set* of rank vectors, we are able to determine more accurately which pages are truly the most important with respect to a particular query or query-context. Because the link-based computations are performed offline, during the preprocessing stage, the query-time costs are not much greater than that of the ordinary PageRank algorithm.

2. REVIEW OF PAGERANK

A review of the PageRank algorithm ([16, 7, 11]) follows. The basic idea of PageRank is that if page u has a link to page v, then the author of u is implicitly conferring some importance to page v. Intuitively, *Yahoo!* is an important page, reflected by the fact that many pages point to it. Likewise, pages prominently pointed to from *Yahoo!* are themselves probably important. How much importance does a page u confer to its outlinks? Let N_u be the outdegree of page u, and let $Rank(p)$ represent the importance (i.e., PageRank) of page p. Then the link (u, v) confers $Rank(u)/N_u$ units of rank to v. This simple idea leads to the following fixpoint computation that yields the rank vector $\vec{Rank}^*$ over all of the pages on the Web. If N is the number of pages, assign all pages the initial value $1/N$. Let B_v represent the set of pages pointing to v. In each iteration, propagate the ranks as follows:[1]

$$\forall_v Rank_{i+1}(v) = \sum_{u \in B_v} Rank_i(u)/N_u \tag{1}$$

We continue the iterations until $\vec{Rank}$ stabilizes to within some threshold. The final vector $\vec{Rank}^*$ contains the PageRank vector over the Web. This vector is computed only once after each crawl of the Web; the values can then be used to influence the ranking of search results [1].

The process can also be expressed as the following eigenvector calculation, providing useful insight into PageRank. Let M be the square, stochastic matrix corresponding to the directed graph G of the Web, assuming all nodes in G have at least one outgoing edge. If there is a link from page j to page i, then let the matrix entry m_{ij} have the value $1/N_j$. Let all other entries have the value 0. One iteration of the previous fixpoint computation corresponds to the matrix-vector multiplication $M \times \vec{Rank}$. Repeatedly multiplying $\vec{Rank}$ by M yields the dominant eigenvector $\vec{Rank}^*$ of the matrix M. In other words, $\vec{Rank}^*$ is the solution to

$$\vec{Rank} = M \times \vec{Rank} \tag{2}$$

Because M corresponds to the stochastic transition matrix over the graph G, PageRank can be viewed as the stationary probability distribution over pages induced by a random walk on the Web.

One caveat is that the convergence of PageRank is guaranteed only if M is irreducible (i.e., G is strongly connected) and aperiodic [15]. The latter is guaranteed in practice for the Web, while the former is true if we add a damping factor $1 - \alpha$ to the rank propagation. We can define a new matrix M' in which we add transition edges of probability $\frac{\alpha}{N}$ between every pair of nodes in G:

$$M' = (1 - \alpha)M + \alpha[\frac{1}{N}]_{N \times N} \tag{3}$$

This modification improves the quality of PageRank by introducing a decay factor $1 - \alpha$ which limits the effect of rank *sinks* [6], in addition to guaranteeing convergence to a unique rank vector. Substituting M' for M in Equation 2, we can express PageRank as the solution to:[2]

$$\vec{Rank} = M' \times \vec{Rank} \tag{4}$$

$$= (1 - \alpha)M \times \vec{Rank} + \alpha\vec{p} \tag{5}$$

with $\vec{p} = [\frac{1}{N}]_{N \times 1}$. The key to creating topic-sensitive PageRank is that we can bias the computation to increase the

[1] Note that for $u \in B_v$, the edge (u, v) guarantees $N_u \geq 1$.

[2] Equation 5 makes use of the fact that $||\vec{Rank}||_1 = 1$.

effect of certain categories of pages by using a nonuniform $N \times 1$ personalization vector for $\vec{p}$ ([6]).[3] Note that the biasing involves introducing additional rank to the appropriate nodes in *each* iteration of the computation. It is not simply a postprocessing step performed on the standard PageRank vector.

In terms of the random-walk model, the personalization vector represents the addition of a complete set of transition edges where the probability on an artificial edge (u, v) is given by αp_v. We will refer to the solution $\vec{Rank}^*$ of Equation 5, with $\alpha = \alpha^*$ and a particular $\vec{p} = \vec{p^*}$, as $\vec{PR}(\alpha^*, \vec{p^*})$. By appropriately selecting $\vec{p}$, the rank vector can be made to prefer certain categories of pages. The bias factor α specifies the degree to which the computation is biased towards $\vec{p}$.

3. TOPIC-SENSITIVE PAGERANK

3.1 Outline of Approach

In our approach to topic-sensitive PageRank, we precompute the importance scores offline, as with ordinary PageRank. However, we compute multiple importance scores for each page; we compute a set of scores of the importance of a page with respect to various topics. At query time, these importance scores are combined based on the topics of the query to form a composite PageRank score for those pages matching the query. This score can be used in conjunction with other IR-based scoring schemes to produce a final rank for the result pages with respect to the query. As the scoring functions of commercial search engines are not known, in our work we do not consider the effect of these other IR scores.[4] We believe that the improvements to PageRank's precision will translate into improvements in overall search rankings, even after other IR-based scores are factored in.[5]

3.2 ODP-biasing

The first step in our approach is to generate a set of biased PageRank vectors using a set of "basis" topics. This step is performed once, offline, during the preprocessing of the Web crawl. For the personalization vector $\vec{p}$ described in Section 2, we use the URLs present in the various categories in the ODP. We create 16 different biased PageRank vectors by using the URLs present below each of the 16 top-level categories of the ODP as the personalization vectors. In particular, let T_j be the set of URLs in the ODP category c_j. Then when computing the PageRank vector for topic c_j, in place of the uniform damping vector $\vec{p} = [\frac{1}{N}]_{N \times 1}$, we use the nonuniform vector $\vec{p} = \vec{v_j}$ where

$$v_{ji} = \begin{cases} \frac{1}{|T_j|} & i \in T_j, \\ 0 & i \notin T_j. \end{cases} \tag{6}$$

The PageRank vector for topic c_j will be referred to as $\vec{PR}(\alpha, \vec{v_j})$. We also generate the single unbiased PageRank vector (denoted as NoBias) for the purpose of comparison. The choice of α will be discussed in Section 4.1.

We also compute the 16 class term-vectors $\vec{D_j}$ consisting of the terms in the documents below each of the 16 top-level categories. D_{jt} simply gives the total number of occurrences of term t in documents listed below class c_j of the ODP.

One could envision using other sources for creating topic-sensitive PageRank vectors; however, the ODP data is freely available, and as it is compiled by thousands of volunteer editors, is less susceptible to influence by any one party.[6]

3.3 Query-Time Importance Score

The second step in our approach is performed at query time. Given a query q, let q' be the context of q. In other words, if the query was issued by highlighting the term q in some Web page u, then q' consists of the terms in u. For ordinary queries not done in context, let $q' = q$. Using a unigram language model, with parameters set to their maximum-likelihood estimates, we compute the class probabilities for each of the 16 top-level ODP classes, conditioned on q'. Let q'_i be the ith term in the query (or query context) q'. Then given the query q, we compute for each c_j the following:

$$P(c_j|q') = \frac{P(c_j) \cdot P(q'|c_j)}{P(q')} \propto P(c_j) \cdot \prod_i P(q'_i|c_j) \tag{7}$$

$P(q'_i|c_j)$ is easily computed from the class term-vector $\vec{D_j}$. The quantity $P(c_j)$ is not as straightforward. We chose to make it uniform, although we could personalize the query results for different *users* by varying this distribution. In other words, for some user k, we can use a prior distribution $P_k(c_j)$ that reflects the interests of user k. This method provides an alternative framework for user-based personalization, rather than directly varying the damping vector $\vec{p}$ as had been suggested in [7, 6].

Using a text index, we retrieve URLs for all documents containing the *original* query terms q. Finally, we compute the query-sensitive importance score of each of these retrieved URLs as follows. Let $rank_{jd}$ be the rank of document d given by the rank vector $\vec{PR}(\alpha, \vec{v_j})$ (i.e., the rank vector for topic c_j). For the Web document d, we compute the query-sensitive importance score s_{qd} as follows.

$$s_{qd} = \sum_j P(c_j|q') \cdot rank_{jd} \tag{8}$$

The results are ranked according to this composite score s_{qd}.[7]

The above query-sensitive PageRank computation has the following probabilistic interpretation, in terms of the "random surfer" model [7]. Let w_j be the coefficient used to weight the jth rank vector, with $\sum_j w_j = 1$ (e.g., let $w_j = P(c_j|q)$). Then note that the equality

$$\sum_j [w_j \vec{PR}(\alpha, \vec{v_j})] = \vec{PR}(\alpha, \sum_j [w_j \vec{v_j}]) \tag{9}$$

[3] A minor caveat: to ensure that M' is irreducible when $\vec{p}$ contains any 0 entries, nodes not reachable from nonzero nodes in $\vec{p}$ should be removed. In practice this is not problematic.

[4] For instance, most search engines use term weighting schemes which make special use of HTML tags.

[5] Note that the topic-sensitive PageRank score itself implicitly makes use of IR in determining the topic of the *query*. However this use of IR is not vulnerable to manipulation of *pages* by adversarial webmasters seeking to raise the score of their sites.

[6] See Section 6 for an approach we are exploring which reduces the ability for even malicious ODP editors to affect scores in any non-negligible way.

[7] Alternatively, s_{qd} can be used as part of a more general scoring function.

holds, as shown in Appendix A. Thus we see that the following random walk on the Web yields the topic-sensitive score s_{qd}. With probability $1-\alpha$, a random surfer on page u follows an outlink of u (where the particular outlink is chosen uniformly at random). With probability $\alpha P(c_j|q')$, the surfer instead jumps to one of the pages in T_j (where the particular page in T_j is chosen uniformly at random). The long term visit probability that the surfer is at page v is exactly given by the composite score s_{qd} defined above. Thus, topics exert influence over the final score in proportion to their affinity with the query (or query context).

4. EXPERIMENTAL RESULTS

To measure the behavior of topic-sensitive PageRank, we conducted a series of experiments. In Section 4.1 we describe the similarity measure we use to compare two rankings. In Section 4.2, we investigate how the induced rankings vary, based on both the topic used to bias the rank vectors as well as the choice of the bias factor α. In Section 4.3, we present results of a user study showing the retrieval performance of ordinary PageRank versus topic-sensitive PageRank. Finally, in Section 4.4, we provide an initial look at how the use of query context can be used in conjunction with topic-sensitive PageRank.

As a source of Web data, we used the latest Web crawl from the Stanford WebBase [12], performed in January 2001, containing roughly 120 million pages. Our crawl contained roughly 280,000 of the 3 million URLs in the ODP. For our experiments, we used 35 of the sample queries given in [9], which were in turn compiled from earlier papers.[8] The queries are listed in Table 1.

Table 1: Queries used

affirmative action	lipari
alcoholism	lyme disease
amusement parks	mutual funds
architecture	national parks
bicycling	parallel architecture
blues	recycling cans
cheese	rock climbing
citrus groves	san francisco
classical guitar	shakespeare
computer vision	stamp collecting
cruises	sushi
death valley	table tennis
field hockey	telecommuting
gardening	vintage cars
graphic design	volcano
gulf war	zen buddhism
hiv	zener
java	

4.1 Similarity Measure for Induced Rankings

We use two measures when comparing rankings. The first measure, denoted $OSim(\tau_1, \tau_2)$, indicates the degree of overlap between the top n URLs of two rankings, τ_1 and τ_2. We define the overlap of two sets A and B (each of size n) to be $\frac{|A \cap B|}{n}$. In our comparisons we will use $n = 20$. The overlap measure $OSim$ gives an incomplete picture of the similarity of two rankings, as it does not indicate the degree to which the relative orderings of the top n URLs of two rankings are in agreement. Therefore, we also use a variant of the Kendall's τ distance measure. See [9] for a discussion of various distance measures for ranked lists in the context of Web search results. For consistency with $OSim$, we will present our definition as a similarity (as opposed to distance) measure, so that values closer to 1 indicate closer agreement. Consider two partially ordered lists of URLs, τ_1 and τ_2, each of length n. Let U be the union of the URLs in τ_1 and τ_2. If δ_1 is $U - \tau_1$, then let τ_1' be the extension of τ_1, where τ_1' contains δ_1 appearing after all the URLs in τ_1.[9] We extend τ_2 analogously to yield τ_2'. We define our similarity measure $KSim$ as follows:

$$KSim(\tau_1, \tau_2) = \frac{|(u,v) : \tau_1', \tau_2' \text{ agree on order of } (u,v), u \neq v|}{|U||U-1|} \tag{10}$$

In other words, $KSim(\tau_1, \tau_2)$ is the probability that τ_1' and τ_2' agree on the relative ordering of a randomly selected pair of distinct nodes $(u, v) \in U \times U$.

4.2 Effect of ODP-Biasing

In this section we measure the effects of topically biasing the PageRank computation. Firstly, note that the choice of the bias factor α, discussed in Section 2, affects the degree to which the resultant vector is biased towards the topic vector used for $\vec{p}$. Consider the extreme cases. For $\alpha = 1$, the URLs in the bias set T_j will be assigned the score $\frac{1}{|T|}$, and all other URLs receive the score 0. Conversely, as α tends to 0, the content of T_j becomes irrelevant to the final score assignment.

We chose to use $\alpha = 0.25$ heuristically, after inspecting the rankings for several of the queries listed in Table 1. We did not concentrate on optimizing α, as we discovered that the induced rankings of query results are not very sensitive to the choice of α. For instance, for $\alpha = 0.05$ and $\alpha = 0.25$, we measured the average similarity of the induced rankings across our set of test queries, for each of our PageRank vectors.[10] The results are given in Table 2. We see that the average overlap between the top 20 results for the two values of α is very high. Furthermore, the high values for $KSim$ indicate high overlap as well agreement (on average) on the relative ordering of these top 20 URLs for the two values of α. All subsequent experiments use $\alpha = 0.25$.

The differences *across* different topically-biased PageRank vectors is much higher, dwarfing any variations caused by the choice of α. We computed the average, across our test queries, of the pairwise similarity between the rankings induced by the different topically-biased vectors. The 5 most similar pairs, according to our $OSim$ measure, are given in Table 3, showing that even the most similar topically-biased rankings have little overlap. Table 4 shows that the pairwise similarities of the rankings induced by the other ranking vectors are close to 0. Having established that the topically-biased PageRank vectors each rank the results substantially differently, we proceed to investigate which of these rankings is "best" for specific queries.

As an example, Table 5 shows the top 5 ranked URLs

[8]Several queries which produced very few hits on our repository were excluded.

[9]The URLs *within* δ_1 are not ordered with respect to one another.

[10]We used 25 iterations of PageRank in all cases.

Table 4: Pairwise comparison of topically-biased rankings ($KSim$)

	NoBias	Arts	Business	Computers	Games	Health	Home	Kids & Teens	News	Recreation	Reference	Regional	Science	Shopping	Society	Sports	World
NoBias	1																
Arts	0.09	1															
Business	0.08	0.06	1														
Computers	0.10	0.08	0.08	1													
Games	0.07	0.12	0.08	0.11	1												
Health	0.07	0.07	0.08	0.06	0.09	1											
Home	0.07	0.07	0.07	0.06	0.09	0.12	1										
Kids & Teens	0.08	0.08	0.04	0.06	0.09	0.11	0.09	1									
News	0.07	0.09	0.07	0.07	0.11	0.09	0.07	0.09	1								
Recreation	0.09	0.09	0.06	0.08	0.09	0.06	0.08	0.08	0.06	1							
Reference	0.07	0.07	0.05	0.08	0.08	0.09	0.06	0.10	0.06	0.05	1						
Regional	0.12	0.09	0.07	0.06	0.06	0.08	0.08	0.08	0.07	0.10	0.07	1					
Science	0.11	0.08	0.08	0.07	0.09	0.11	0.06	0.09	0.08	0.06	0.10	0.08	1				
Shopping	0.05	0.07	0.07	0.06	0.09	0.06	0.07	0.05	0.05	0.08	0.04	0.06	0.04	1			
Society	0.10	0.10	0.06	0.06	0.07	0.10	0.09	0.11	0.09	0.08	0.09	0.11	0.10	0.05	1		
Sports	0.07	0.09	0.07	0.07	0.13	0.09	0.10	0.08	0.10	0.10	0.07	0.09	0.07	0.09	0.07	1	
World	0.10	0.06	0.06	0.07	0.07	0.06	0.05	0.06	0.06	0.07	0.06	0.08	0.07	0.05	0.07	0.06	1

Table 2: Average similarity of rankings for $\alpha = \{0.05, 0.25\}$

Bias Set	$OSim$	$KSim$
NoBias	0.72	0.64
Arts	0.66	0.58
Business	0.63	0.54
Computers	0.70	0.60
Games	0.78	0.67
Health	0.73	0.62
Home	0.77	0.67
Kids & Teens	0.74	0.66
News	0.74	0.65
Recreation	0.62	0.55
Reference	0.68	0.57
Regional	0.60	0.52
Science	0.69	0.59
Shopping	0.66	0.55
Society	0.57	0.50
Sports	0.69	0.60
World	0.64	0.55

Table 3: Topic pairs yielding most similar rankings

Bias-Topic Pair	$OSim$	$KSim$
(Games, Sports)	0.18	0.13
(NoBias, Regional)	0.18	0.12
(Kids & Teens, Society)	0.18	0.11
(Health, Home)	0.17	0.12
(Health, Kids & Teens)	0.17	0.11

for the query "bicycling," using each of the topically-biased PageRank vectors. Note in particular that the ranking induced by the Sports-biased vector is of high quality.[11] Also note that the ranking induced by the Shopping-biased vector leads to the high ranking of websites selling bicycle-related accessories.

4.3 Query-Sensitive Scoring

In this section we look at how effectively we can utilize the ranking precision gained by the use of multiple PageRank vectors. Given a query, our first task is to determine which of the rank vectors can best rank the results for the query. We found that using the quantity $P(c_j|q)$ as discussed in Section 3.3 yielded intuitive results for determining which topics are most closely associated with a query. In particular, for most of the test queries, the ODP categories with the highest values for $P(c_j|q)$ are intuitively the most relevant categories for the query. In Table 6, we list for each test query, the 3 categories with the highest values for $P(c_j|q)$. When computing the composite s_{qd} score in our experiments, we chose to use the weighted sum of only the rank vectors associated with the three topics with the highest values for $P(c_j|q)$, rather than all of the topics. Based on the data in Table 6, we saw no need to include the scores from the topic vectors with lower associated values for $P(c_j|q)$.

To compare our query-sensitive approach to ordinary PageRank, we conducted a user study. We randomly selected 10 queries from our test set for the study, and found 5 volunteers. For each query, the volunteer was shown 2 result rankings; one consisted of the top 10 results satisfying the query, when these results were ranked with the unbiased PageRank vector, and the other consisted of the top 10 results for the query when the results were ranked with the composite s_{qd} score.[12] The volunteer was asked to select all URLs which were "relevant" to the query, in their opinion. Furthermore, they were asked to say which of the two

[11]Of course this is a subjective statement; a user study is presented in Section 4.3.

[12]Both the title and URL were presented to the user. The title was a hyperlink to a current version of the Web page.

Table 5: Query “bicycling”

NoBias “RailRiders Adventure Clothing” www.RailRiders.com www.Waypoint.org/default.html www.Gorp.com/ www.FloridaCycling.com/ HiddenTrails.com/index.htm	**Arts** “Photo Contest & Gallery (Bicycling)” www.bikescape.com/photogallery/ www.trygve.com/ www.greenway.org/ www.jsc.nasa.gov/Bios/htmlbios/young.html www.BellaOnline.com/sports/
Business “Recumbent Bikes and Kit Aircraft” www.rans.com www.BreakawayBooks.com java.oreilly.com/bite-size/ www.carbboom.com www.CorporateTeamBuilding.com	**Computers** “GPS Pilot” www.gpspilot.com www.wireless.gr/wireless-links.htm www.linkstosales.com www.LiftExperts.com/lifts.html www.trygve.com/index.html
Games “Definition Through Hobbies” www.flick.com/~gretchen/hobbies.html www.BellaOnline.com/sports/ www.npr.org/programs/wesun/puzzle/will.html www.trygve.com/ www.IdeaFinder.com/showcase/forsale.htm	**Health** “Personal Fitness Trainer...” www.nfpt.com/guestbook.html www.usrf.org/news/bikeriding.html obgyn.uihc.uiowa.edu/Patinfo/Adhealth/UTI.HTM www.nmh.org/ www.ChainreactionBicycles.com/saddles.htm
Home “25 Ways to Stay On Track” www.exercare.com/exerinfo/motivation.htm www.floras-hideout.com/party/index.html www.BicycleSource.com/contact.shtml www.bicycleSource.com/tour.shtml www.aoa.dhhs.gov/elderpage.html	**Kids and Teens** “Camp Shohola For Boys” www.shohola.com www.EarthForce.org www.WeissmanTours.com www.GrownupCamps.com/homepage.html www.EarthForce.org/welcome.htm
News “Minnesotans for an Energy-Efficient Economy” www.me3.org/projects/sprawl/ www.SmithfieldTimes.com/TimesEditorl.htm www.DaveSloan.com/about/ www.TheAtlantic.com/issues/2000/11/russo.htm www.SierraClub.org/ico/	**Recreation** “Adventure travel” www.gorp.com/ www.GrownupCamps.com/homepage.html www.gorp.com/gorp/activity/main.htm www.outdoor-pursuits.org/ www.NicholsExpeditions.com/
Reference “WPI Clubs & Organizations” www.wpi.edu/Admin/SAO/guide.html www.NoyesFamily.com/school/manciano.html www.ThePotters.com/puzzles.html www.Vanderbilt.edu/AnS/Germanic-Slavic/german/ www.engin.umich.edu/prog/macro/univA2.html	**Regional** “Your Guide to Outdoor Activities” www.gorp.com/gorp/activity/main.htm www.gorp.com/ www.destateparks.com/index.htm www.tpwd.state.tx.us/park/parks.htm www.gorp.com/gorp/activity/biking.htm
Science “Coast to Coast by Recumbent Bicycle” hypertextbook.com/bent/ www.SiestaSoftware.com/ www.BenWiens.com/benwiens.html www.SusanJeffers.com/jeffbio.htm www.EarthForce.org/welcomeA.htm	**Shopping** “Cycling Clothing & Accessories for Women” www.TeamEstrogen.com/ www.ShopOutdoors.com/ www.jub.com.au/books/ www.bike.com/ www.softride.com/
Society “Word Search Puzzles” www.ThePotters.com/puzzles.html www.LakeTravisbb.com/ www.vnorthland.com/hotel/barkpoint/barkpoint.htm www.gorp.com/default.htm www.tlcnetwork.org/	**Sports** “Swim, Bike, Run, & Multisport” www.multisports.com/ www.BikeRacing.com/ www.CycleCanada.com/ www.bikescape.com/photogallery/ www.cambiecycles.com/
World “Disease Word Index” www.pathinfo.com/lhodzpds.htm www.ExploringEcuador.com/espindex.htm www.camembert-france.com/bike00.html www.AdventureRace.com/JungleMan.htm www.dejava.com/yogya/transits.htm	

Table 7: Ranking preferred by majority of users

Query	Preferred by Majority
alcoholism	TopicSensitive
bicycling	TopicSensitive
citrus groves	TopicSensitive
computer vision	TopicSensitive
death valley	TopicSensitive
graphic design	TopicSensitive
gulf war	TopicSensitive
hiv	NoBias
shakespeare	Neither
table tennis	TopicSensitive

rankings was "better" overall, in their opinion. They were not told anything about how either of the rankings was generated. The rankings induced by the topic-sensitive PageRank score s_{qd} were significantly preferred by our test group. Let a URL be considered *relevant* if at least 3 of the 5 volunteers selected it as relevant for the query. The *precision* then is the fraction of the top 10 URLs that are deemed *relevant*. The precision of the two ranking techniques for each test query is shown in Figure 1. The average precision for the rankings induced by the topic-sensitive PageRank scores is substantially higher than that of the unbiased PageRank scores. Furthermore, as shown in Table 7, for nearly all queries, a majority of the users preferred the rankings induced by the topic-sensitive PageRank scores. These results suggest that the effectiveness of a query-result scoring function can be improved by the use of a topic-sensitive PageRank scheme in place of a generic PageRank scheme.

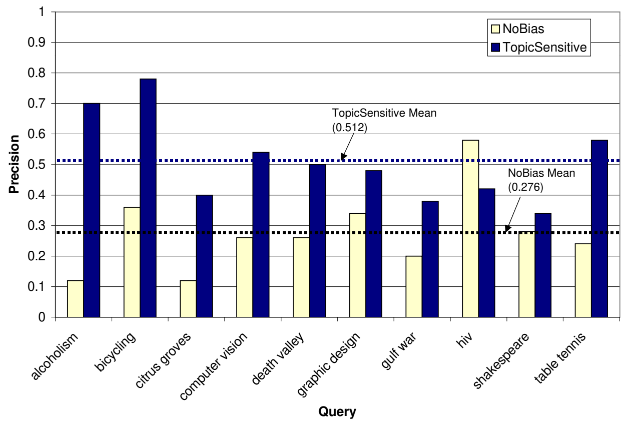

Figure 1: Precision @ 10 results for our test queries. The average precision over the ten queries is also shown.

4.4 Context-Sensitive Scoring

In Section 4.3, the topic-sensitive ranking vectors were chosen using the topics most strongly associated with the query term. If the search is done in context, for instance by highlighting a term in a Web page and invoking a search, then the context can be used instead of the query to determine the topics. Using the context can help disambiguate the query term and yield results that more closely reflect the intent of the user. We now illustrate with an example how using query-context can help a system which uses topic-sensitive PageRank.

Consider the query "blues" taken from our test set. This term has several different senses; for instance it could refer to a musical genre, or to a form of depression. Two Web pages in which the term is used with these different senses, as well as short textual excerpts from the pages, are shown in Table 8. Consider the case where a user reading one of these two pages highlights the term "blues" to submit a search query. At query time, the first step of our system is to determine which topic best applies to the query in context. Thus, we calculate $P(c_j|q')$ as described in Section 3.3, using for q' the terms of the entire page, rather than just the term "blues." For the first page (discussing music), $argmax_{c_j} P(c_j|q')$ is Arts, and for the second page (discussing depression), $argmax_{c_j} P(c_j|q')$ is Health. The next step is to use a text index to fetch a list of URLs for all documents containing the term "blues" — the highlighted term for which the query was issued. Finally, the URLs are ranked using the appropriate ranking vector that was selected using the $P(c_j|q')$ values (i.e., either Arts or Health). Table 9 shows the top 5 URLs for the query "blues" using the topic-sensitive PageRank vectors for Arts, Health, and NoBias. We see that as desired, most of the results ranked using the Arts-biased vector are pages discussing music, while all of the top results ranked using the Health-biased vector discuss depression. The context of the query allows the system to pick the appropriate topic-sensitive ranking vector, and yields search results reflecting the appropriate sense of the search term.

5. SOURCES OF SEARCH CONTEXT

In the previous section, we discussed one possible source of context to utilize in the generation of the composite PageRank score, namely the document containing the query term highlighted by the user. There are a variety of other sources of context that may be used in our scheme. For instance, the history of queries issued leading up to the current query is another form of query context. A search for "basketball" followed up with a search for "Jordan" presents an opportunity for disambiguating the latter. As another example, most modern search engines incorporate some sort of hierarchical directory, listing URLs for a small subset of the Web, as part of their search interface.[13] The current node in the hierarchy that the user is browsing at constitutes a source of query context. When browsing URLs at Top/Arts, for instance, any queries issued could have search results (from the entire Web index) ranked with the Arts rank vector, rather than either restricting results to URLs listed in that particular category, or not making use of the category whatsoever. In addition to these types of context associated with the query itself, we can also potentially utilize query independent *user* context. Sources of user context include the users' browsing patterns, bookmarks, and email archives. As mentioned in Section 3.3, we can integrate user context by selecting a nonuniform prior, $P_k(c_j)$, based on how closely the user's context accords with each of the basis topics.

When attempting to utilize the aforementioned sources of search context, mediating the personalization of PageRank

[13] See for instance `http://directory.google.com/Top/Arts/` or `http://dir.yahoo.com/Arts/`.

Table 6: Estimates for $P(c_j|q)$

affirmative action	
News	0.41
Society	0.22
Reference	0.17

alcoholism	
Health	0.47
Kids & Teens	0.20
Arts	0.06

amusement parks	
Regional	0.51
Recreation	0.23
Kids & Teens	0.08

architecture	
Computers	0.26
Reference	0.19
Business	0.09

bicycling	
Sports	0.52
Regional	0.13
Health	0.07

blues	
Arts	0.52
Shopping	0.12
News	0.08

cheese	
Home	0.72
Recreation	0.10
Shopping	0.05

citrus groves	
Shopping	0.34
Home	0.21
Regional	0.18

classical guitar	
Arts	0.75
Shopping	0.21
News	0.01

computer vision	
Computers	0.24
Business	0.14
Reference	0.09

cruises	
Recreation	0.65
Regional	0.18
Sports	0.04

death valley	
Regional	0.28
Society	0.14
News	0.10

field hockey	
Sports	0.89
Shopping	0.03
Reference	0.03

gardening	
Home	0.63
Shopping	0.14
Regional	0.04

graphic design	
Computers	0.36
Business	0.23
Shopping	0.09

gulf war	
Society	0.21
Kids & Teens	0.18
Regional	0.17

hiv	
Health	0.40
News	0.19
Kids & Teens	0.14

java	
Computers	0.53
Games	0.10
Kids & Teens	0.06

lipari	
Home	0.19
Kids & Teens	0.17
News	0.13

lyme disease	
Health	0.96
Regional	0.01
Recreation	0.01

mutual funds	
Business	0.77
Regional	0.05
Home	0.05

national parks	
Regional	0.42
Recreation	0.16
Kids & Teens	0.09

parallel architecture	
Computers	0.70
Science	0.10
Reference	0.07

recycling cans	
Home	0.42
Business	0.38
Kids & Teens	0.06

rock climbing	
Recreation	0.54
Regional	0.13
Sports	0.07

san francisco	
Sports	0.27
Regional	0.16
Recreation	0.10

shakespeare	
Arts	0.34
Reference	0.21
Kids & Teens	0.15

stamp collecting	
Shopping	0.44
Recreation	0.39
Science	0.02

sushi	
Home	0.56
Kids & Teens	0.13
Shopping	0.07

table tennis	
Sports	0.53
Shopping	0.14
Regional	0.09

telecommuting	
Business	0.70
Kids & Teens	0.04
Society	0.03

vintage cars	
Shopping	0.67
Recreation	0.23
Home	0.02

volcano	
Science	0.36
Regional	0.18
Recreation	0.13

zen buddhism	
Society	0.88
Kids & Teens	0.09
World	0.01

zener	
Kids & Teens	0.17
News	0.13
Business	0.11

Table 8: Two different search contexts for the query "blues"

That Blues Music Page	Postpartum Depression & the 'Baby Blues'
http://www.fred.net/turtle/blues.shtml	http://familydoctor.org/handouts/379.html
... If you're stuck for new material, visit Dan Bowden's Blues and Jazz Transcriptions - lots of older blues guitar transcriptions for you historic blues fans ...	... If you're a new mother and have any of these symptoms, you have what is called the "baby blues." "The blues" are considered a normal part of early motherhood and usually go away within 10 days after delivery. However, some women have worse symptoms or symptoms last longer. This is called "postpartum depression." ...

Table 9: Results for query "blues"

Arts
Britannica Online www.britannica.com
BandHunt.com Genres (Music) www.bandhunt.com/genres.html
Artist Information (Music) www.artistinformation.com/index.html
Billboard.com (Music charts) www.billboard.com
Soul Patrol (Music) www.soul-patrol.com

Health
Northern County Psychiatric Associates News www.baltimorepsych.com/news.htm
Seasonal Affective Disorder www.ncpamd.com/seasonal.htm
Women's Mental Health www.ncpamd.com/Women's_Mental_Health.htm
Wing of Madness Depression Support Group www.wingofmadness.com
Country Nurse Online www.countrynurse.com

NoBias
TUCOWS Themes news.tucows.com/themes/pastart.html
World's Most Popular MP3 Service www.emusic.com
Books, Music, DVD, and VHS Essentials www.johnholleman.com/amastatement.html
The Official Site of Major League Baseball www.majorleaguebaseball.com
MP3.com: Free MP3 Downloads www.mp3.com

via a set of basis topics yields several benefits over attempting to explicitly choose a personalization vector directly.

Flexibility: For any kind of context, we can compute the context-sensitive PageRank score by using a classifier to compute the similarity of the context with the basis topics and then weighting the topic-sensitive PageRank vectors appropriately. We can treat such diverse sources of search context such as email, bookmarks, browsing history, and query history uniformly.

Transparency: The topically-biased rank vectors have intuitive interpretations. If we see that our system is giving undue preference to certain topics, we can tune the classifier used on the search context, or adjust topic weights manually. When utilizing user context, the users themselves can be shown what topics the system believes represent their interests.

Privacy: Certain forms of search context raise potential privacy concerns. Clearly it is inappropriate to send the user's browsing history or other personal information to the search-engine server for use in constructing a profile. However a *client-side* program could use the user context to generate the user profile locally, and send only the summary information, consisting of the weights assigned to the basis topics, over to the server. The amount of privacy lost in knowing only that the user's browsing pattern suggests that he is interested in COMPUTERS with weight 0.5 is much less than actually obtaining his browser's history cache. When making use of query-context, if the user is browsing sensitive personal documents, they would be more comfortable if the search client sent to the server topic weights rather than the actual document text surrounding the highlighted query term.

Efficiency: For a small number of basis topics (such as the 16 ODP categories), both the query-time cost and the offline preprocessing cost of our approach is low, and practical to implement with current Web indexing infrastructure.

A wide variety search-context sources exist which, if utilized appropriately, can help users better manage the deluge of information they are faced with. Although we have begun exploring how best to make use of available context, much work remains in identifying and utilizing search context with the goal of personalizing Web search.

6. ONGOING WORK

We are currently exploring several ways of improving our approach for topic-sensitive PageRank. As discussed in the previous section, discovering sources of search context is a ripe area of research. Another area of investigation is the development of the best set of basis topics. For instance it may be worthwhile to use a finer-grained set of topics, perhaps using the second or third level of the Open Directory hierarchy, rather than simply the top level. However, a fine-grained set of topics leads to efficiency considerations, as the cost of the naive approach to computing these topic-sensitive vectors is linear in the number of basis topics. See [13] for approaches that may make the use of a larger, finer grained set of basis topics practical.

We are also currently investigating a different approach to creating the damping vector $\vec{p}$ used to create the topic-sensitive rank vectors. This approach has the potential of being more resistant to adversarial ODP editors. Currently, as described in Section 3.2, we set the damping vector $\vec{p}$ for topic c_j to $\vec{v_j}$, where $\vec{v_j}$ is defined in Equation 6. In the modified approach, we instead first train a classifier for the basis set of topics using the ODP data as our training set, and then assign to *all* pages on the Web a distribution of topic weights.[14] Let this topic weight of a page u for category c_j be w_{uj}. Then we replace Equation 6 with

$$\forall_{i \in Web}[v_{ji} = \frac{w_{ij}}{\sum_k w_{kj}}] \tag{11}$$

In this way, we hope to ensure that the PageRank vectors generated are not overly sensitive to particular choices made

[14]For instance, the estimated class probabilities for the basis topics.

by individual ODP editors.

We plan to investigate the above enhancements to generating the topic-sensitive PageRank score, and evaluate their effect on retrieval performance, both in isolation and when combined with typical IR scoring functions.

7. ACKNOWLEDGMENTS

I would like to thank Professor Jeff Ullman for invaluable comments and feedback. I would like to thank Glen Jeh and Professor Jennifer Widom for several useful discussions. I would also like to thank Aristides Gionis for his feedback. Finally, I would like to thank the anonymous reviewers for their insightful comments.

8. REFERENCES

[1] The Google Search Engine: Commercial search engine founded by the originators of PageRank. http://www.google.com/.

[2] The Open Directory Project: Web directory for over 2.5 million URLs. http://www.dmoz.org/.

[3] 'More Evil Than Dr. Evil?' http://searchenginewatch.com/sereport/99/11-google.html.

[4] Krishna Bharat and Monika R. Henzinger. Improved algorithms for topic distillation in a hyperlinked environment. In *Proceedings of the ACM-SIGIR*, 1998.

[5] Krishna Bharat and George A. Mihaila. When experts agree: Using non-affiliated experts to rank popular topics. In *Proceedings of the Tenth International World Wide Web Conference*, 2001.

[6] Sergey Brin, Rajeev Motwani, Larry Page, and Terry Winograd. What can you do with a web in your pocket. In *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 1998.

[7] Sergey Brin and Larry Page. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the Seventh International World Wide Web Conference*, 1998.

[8] S. Chakrabarti, B. Dom, D. Gibson, J. Kleinberg, P. Raghavan, and S. Rajagopalan. Automatic resource compilation by analyzing hyperlink structure and associated text. In *Proceedings of the Seventh International World Wide Web Conference*, 1998.

[9] Cynthia Dwork, Ravi Kumar, Moni Naor, and D. Sivakumar. Rank aggregation methods for the web. In *Proceedings of the Tenth International World Wide Web Conference*, 2001.

[10] Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. Placing search in context: the concept revisited. In *Proceedings of the Tenth International World Wide Web Conference*, 2001.

[11] Taher H. Haveliwala. Efficient computation of PageRank. *Stanford University Technical Report*, 1999.

[12] J. Hirai, S. Raghavan, H. Garcia-Molina, and A. Paepcke. Webbase: A repository of web pages. In *Proceedings of the Ninth International World Wide Web Conference*, 2000.

[13] Glen Jeh and Jennifer Widom. Scaling personalized web search. *Stanford University Technical Report*, 2002.

[14] Jon Kleinberg. Authoritative sources in a hyperlinked environment. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*, 1998.

[15] Rajeev Motwani and Prabhakar Raghavan. *Randomized Algorithms*. Cambridge University Press, United Kingdom, 1995.

[16] Larry Page. PageRank: Bringing order to the web. *Stanford Digital Libraries Working Paper*, 1997.

[17] Davood Rafiei and Alberto O. Mendelzon. What is this page known for? Computing web page reputations. In *Proceedings of the Ninth International World Wide Web Conference*, 2000.

[18] Matthew Richardson and Pedro Domingos. *The Intelligent Surfer: Probabilistic Combination of Link and Content Information in PageRank*, volume 14. MIT Press, Cambridge, MA, 2002 (To appear).

APPENDIX

A. WEIGHTED SUM OF PAGERANK VECTORS

In this section we derive the interpretation of the weighted sum of PageRank vectors.[15] Consider a set of rank vectors $\{\vec{PR}(\alpha, \vec{v_j})\}$ for some fixed α.[16] For brevity let $\vec{r_j} = \vec{PR}(\alpha, \vec{v_j})$. Furthermore let $\vec{r'} = \sum_j [w_j \vec{r_j}]$, and $\vec{v'} = \sum_j [w_j \vec{v_j}]$. We claim that $\vec{r'} = \vec{PR}(\alpha, \vec{v'})$. In other words, $\vec{r'}$ is itself a PageRank vector, where the personalization vector $\vec{p}$ is set to $\vec{v'}$. The proof follows.

Because each $\vec{r_j}$ satisfies Equation 5 (with $\vec{p} = \vec{v_j}$), we have that

$$\begin{aligned}
\vec{r'} &\equiv \sum_j [w_j \vec{r_j}] && (12)\\
&= \sum_j [w_j ((1-\alpha) M \vec{r_j} + \alpha \vec{v_j})] && (13)\\
&= \sum_j [(1-\alpha) w_j M \vec{r_j}] + \sum_j [\alpha w_j \vec{v_j}] && (14)\\
&= (1-\alpha) M \sum_j [w_j \vec{r_j}] + \alpha \sum_j [w_j \vec{v_j}] && (15)\\
&= (1-\alpha) M \vec{r'} + \alpha \vec{v'} && (16)
\end{aligned}$$

Thus $\vec{r'}$ satisfies Equation 5 for the personalization vector $\vec{p} = \vec{v'}$, and our proof is complete.

[15] The proof that follows is based on discussions with Glen Jeh (see [13]).

[16] See the end of Section 2 for the description of our notation.