Topical Link Analysis for Web Search

Lan Nie   Brian D. Davison   Xiaoguang Qi
Department of Computer Science & Engineering
Lehigh University
Bethlehem, PA 18015 USA
{lan2,davison,xiq204}@cse.lehigh.edu

ABSTRACT

Traditional web link-based ranking schemes use a single score to measure a page's authority without concern of the community from which that authority is derived. As a result, a resource that is highly popular for one topic may dominate the results of another topic in which it is less authoritative. To address this problem, we suggest calculating a score vector for each page to distinguish the contribution from different topics, using a random walk model that probabilistically combines page topic distribution and link structure. We show how to incorporate the topical model within both PageRank and HITS without affecting the overall property and still render insight into topic-level transition. Experiments on multiple datasets indicate that our technique outperforms other ranking approaches that incorporate textual analysis.

Categories and Subject Descriptors: H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval

General Terms: Algorithms, Performance

Keywords: Web search engine, link analysis, HITS, PageRank

1. INTRODUCTION

The use of link analysis techniques to estimate the importance of a page made web search engines useful, and was central to the meteoric rise of Google as the leading commercial engine. Today, however, traditional web link analysis is necessary, but insufficient: Google presently claims more than one hundred factors in calculating the results of a query [7]. One possible reason for this is that in traditional web link analysis, a page's authority is measured by the summation of incoming authority flows, without concern for the community from which that authority is derived. As a result, a resource that is highly popular for one topic may dominate the results of another topic in which it is less authoritative. For example, a popular news website will be ranked highly for the query "Shakespeare" only because a document included a quotation from Shakespeare's writings. Obviously it is unfair for a site's reputation in news to dominate when the context is classic literature. However, with a single generic measurement for reputation, it is impossible to tell them apart and determine for what aspects this page is known.

Trusting someone with some aspect of your life (say, to care for a child) does not mean that the person should also be entrusted in other areas (e.g., to decide how to invest your finances). The same is true for reputation or authority of pages on the Web. Instead, a mechanism is needed to determine a topical context for statements of authority. For web pages, we propose that in the calculation of authority, incoming authority flows should be split across different topics instead of being mixed together indiscriminately. The effect is to separate the authority score into a vector to record a page's reputation with respect to different topics.

There are many potential applications for such computation. First, this separation will avoid the problem of a heavily linked page getting highly ranked for an irrelevant query, thus yielding more accurate search results. A second application is categorization; determining the topic on which a page has highest reputation in the vector is evidence that the page is about that topic. In addition, given a topic, we could rank pages according to their topic-specific authority; the resulting higher ranked pages for a topic could then be selected as good candidates to be included in a directory of resources on the topic. This provides a means of automatic page annotation, which is more flexible and economic than manually organized Web directories such as the dmoz Open Directory Project (ODP) [13] and the Yahoo directory [20].

In this paper, we propose a topical link analysis model that formalizes this intuition by using a topic distribution to embody the context in which a link is created and thus affect the authority propagation. We show how to incorporate the topical model within both PageRank [14, 3] and HITS [11] without affecting the overall authority (or hub) score, and still provide a global analysis (not just a subset) that can be interpreted in a query or topic-specific manner.

The contributions of the paper are:

- A novel method incorporating topical features into PageRank and HITS without affecting their global properties, while providing insight into the topic-level transition within the global authority propagation.
- An extensive experimental comparison of our approach to a number of well-known ranking algorithms to show the superiority of our approach.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
SIGIR'06, August 6–11, 2006, Seattle, Washington, USA.
Copyright 2006 ACM 1-59593-369-7/06/0008 ...$5.00.

The remainder of this paper is organized as follows: the background and related work will be introduced in Section 2, with a focus on combining text and link analysis. The topical link analysis model is then detailed, with the novel generalizations of well-known link analysis techniques and additional issues that affect topical approaches. The experiments and results will be shown in Section 4. We conclude with a discussion and future work.

2. RELATED WORK

Much research has considered the union of text and link analysis and some even consider the issue of topicality. We review them here.

We start by introducing some notation:

- $I(v)$: in-degree of page v
- $O(v)$: out-degree of page v
- $A(v)$: authority score of page v
- $H(v)$: hubness score of page v
- W: the set of web pages
- N: the number of pages in W
- d: the probability of a random jump in the random surfer model
- $p \rightarrow q$: there is a hyperlink on page p that points to q

2.1 Hyperlink-Induced Topic Search (HITS)

While at IBM Research, Jon Kleinberg proposed [11] that web documents had two important properties, called hubness and authority, as well as a mechanism to calculate them. Pages functioning as good hubs have links pointing to many good authority pages, and good authorities are pages to which many good hubs point. Thus, in his Hyperlink-Induced Topic Search (HITS) approach to broad topic information discovery, the score of a hub (authority) depended on the sum of the scores of the connected authorities (hubs):

$$A(p) = \sum_{q:q\rightarrow p} H(q) \text{ and } H(p) = \sum_{q:p\rightarrow q} A(q)$$

Kleinberg didn't calculate these scores globally; instead, he used the subset of the web that included top-ranked pages for a given query, plus those pages pointed to and were pointed by that set, the union of which he called the base set. When first introduced, HITS was able to work with the existing search engines (which were mostly textually based) and generate results (at least for broad topics) that were comparable to hand-compiled information in directories such as Yahoo.

HITS uses the results of a search engine query to make its analysis query-specific, but sometimes the topical focus of the base set can drift to a broader or more popular topic. In addition, the one-step expansion can bring in unrelated yet popular pages that can end up with high ranks. A number of improvements have since been proposed.

CLEVER improvements to HITS. IBM's CLEVER [9, 5] project extended HITS. The ARC algorithm [6] expands the core set with nodes up to two steps away, and weights links by the similarity between the query and the text surrounding the hyperlink.

Bharat and Henzinger's improvements to HITS. Bharat and Henzinger [2] proposed a number of improvements to HITS. The first change is an algorithm called *imp*, which re-weights links involved in mutually reinforcing relationships and drops links within the same host. In order to reduce topic drift, they eliminate documents that were not sufficiently similar to the query topic, which was comprised of the first 1,000 words of each core document. In addition, they used the relevance scores of a node as a weight on its contribution so that the nodes most relevant to the query have the most influence on the calculation. They found that *imp* made a big improvement over the original HITS .

2.2 PageRank

At approximately the same time as Kleinberg, Stanford graduate students Sergey Brin and Lawrence Page proposed an alternative model of page importance, called the random surfer model [14]. In that model, a surfer on a given page i, with probability $(1-d)$ chooses to select uniformly one of its outlinks $O(i)$, and with probability d to jump to a random page from the entire web W. The PageRank [3] score for node i is defined as the stationary probability of finding the random surfer at node i. One formulation of PageRank is

$$PR(i) = (1-d) \sum_{j:j\rightarrow i} \frac{PR(j)}{O(j)} + d\frac{1}{N}$$

Because the definition of PageRank is recursive, it must be iteratively evaluated until convergence.

PageRank is a topic-independent measure of the importance of a web page, and must be combined with one or more measures of query relevance for ranking the results of a search.

Topic-Sensitive PageRank. In Haveliwala's Topic-Sensitive PageRank (TSPR) [8], multiple PageRank calculations are performed, one per topic. Selected topics consisted of the top level categories of the ODP, with τ_j as the set of URLs within topic c_j. When computing the PageRank vector for category c_j, the random surfer will jump to a page in τ_j at random rather than just any web page. This has the effect of biasing the PageRank to that topic. Thus, page k's score on topic c_j can be defined as

$$TSPR_j(k) = (1-d) \sum_{i:i\rightarrow k} \frac{TSPR_j(i)}{O(i)} + \begin{cases} d\frac{1}{|\tau_j|} & \text{if } k \in \tau_j \\ 0 & \text{if } k \notin \tau_j \end{cases}$$

To be able to rank results for a particular query q, let $r(q, c_j)$ be q's relevance to topic c_j. For web page k, the query sensitive importance score is

$$S_q(k) = \sum_j TSPR_j(k) * r(q, c_j)$$

The results are ranked according to this composite score.

While Haveliwala biased the PageRank to a specific topic in the jump session, in contrast, Pal and Narayan [15] adopted this kind of biasing when following a link: instead of selecting among all of the outlinks uniformly, a focus surfer on topic c_j favors those leading to pages on topic c_j, the off-topic pages may be visited occasionally, but with a much smaller probability. However, this model didn't take the surfer's jump behavior into consideration.

The Intelligent Surfer. In Richardson and Domingos' Intelligent Surfer [17], the surfer is prescient, selecting links (or jumps) based on the relevance of the target to the query of interest. In such a query-specific version of PageRank, the surfer still has two choices: follow a link, with probability $(1-d)$, or jump, with probability d. However, instead of selecting among the possible destinations equally, the surfer chooses using a probability distribution generated from the relevance of the target to the surfer's query.

Thus, for a specific query q, page j's query-dependent score can be calculated as

$$IS_q(j) = d\frac{r(q,j)}{\sum_{k\in W} r(q,k)} + (1-d)\sum_{i:i\to j} IS_q(i)\frac{r(q,j)}{\sum_{l:i\to l} r(q,l)}$$

Since it is not feasible to calculate this query-specific PageRank at run-time, term-specific PageRanks are generated in advance for all terms, and in the case of a multi-term query, the resulting page scores are combined using the weights of the terms from the query.

3. TOPICAL LINK ANALYSIS

In this section, we argue that every aspect of reputation has a context within which it is interpreted, and that incorporating that context into automated analyses of reputation can potentially provide more accurate models. In some cases, it can provide features not otherwise possible. This view is broad, and certainly encompasses many of the approaches we have described above.

In the rest of this section, we will start by providing an overview of the topical link analysis approach that we use, and then apply it to the two well-known traditional link analysis frameworks: PageRank and HITS.

3.1 Overview

The basic idea of topical link analysis is to incorporate topic distribution into the representation of each web page as well as the importance score of each page. Therefore, there are at least two vectors associated with each page: the content vector and the authority vector.

The content vector C_u:$[C(u_1), C(u_2), ..., C(u_T)]$ is a probability distribution used to represent the content of u, in which each component represents the the relative contribution from each topic within the content of u to the content of u as a whole. This vector is static and solely determined by the content. We can use a textual classifier to provide such a soft classification for each document (or query) across T predefined topics. As shown in Figure 1, a page's content is represented by the corresponding content vector in the topic level. This vector is normalized such that the sum of the probabilities is 1. Since a query can also be viewed as a short document, query q can also have an associated content vector C_q, which may be generated by a textual classifier, to represent its relevance to each topic.

In contrast, we assign each page u an authority vector A_u:$[A(u_1), A(u_2), ..., A(u_T)]$ to measure its importance, where $A(u_k)$ denotes page u's importance score on topic k. (In the HITS version, besides the authority vector, there is a corresponding hubness vector as well.) This vector is obtained from the proposed topical ranking algorithm, which is dynamic during authority propagation. From Figure 1, we can tell that the summation $A(u) = \sum_{k\in T} A(u_k)$ is identical to the original non-topical importance score, e.g., the score obtained by the PageRank algorithm, and a page's authority distribution may differ from its content distribution in the topic level.

When the authority vector for each page is ready, a query-specific importance score can be calculated for each query using $S_q(u) = \sum_k A(u_k) * C(q_k)$, where the components in the authority vector A_u are weighted by the query's relevance distribution, as given in C_q.

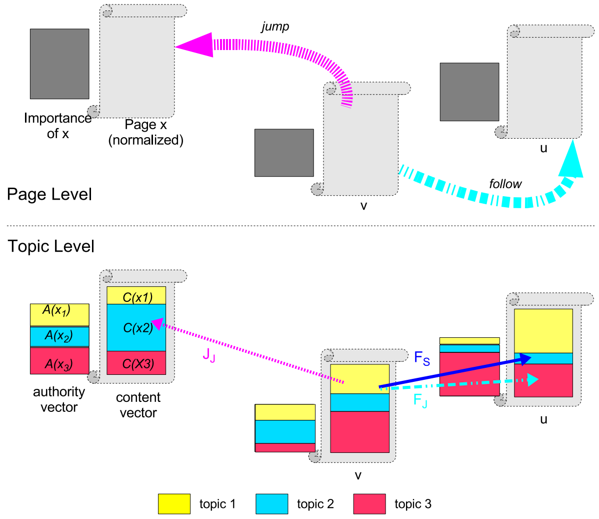


Figure 1: Illustration of topics within nodes.

So far, we presented an overview of the topical link analysis approach. In the following, two models for computing the authority vectors will be proposed: Topical PageRank and Topical HITS.

3.2 Topical PageRank Model

In order to introduce Topical PageRank, we will start by describing the "topical random surfer" model.

A "topical random surfer" is similar to the "random surfer" described in the PageRank model. The difference is that the topical one is sensitive to different topics. Consider a "topical random surfer" who wanders on the Web. Assume the surfer is browsing a web page v for he/she is interested in topic k on that page. For the next move, the surfer may either *follow* an outgoing link on the current page with probability $(1-d)$ or *jump* to any page uniformly at random with probability d. As PageRank usually does, d is set as 0.15 here.

Intuitively, when following a link, the surfer is likely to stay on the same topic to maintain topic continuity (*follow-stay*, "F_S", as shown in Figure 1); however, with a probability $(1-\alpha)$, he/she may jump to any topic i in the target page (*follow-jump*, "F_J"). When taking an "F_J" action, the preference among topics is determined by the content in the target page u, i.e., C_u. In the example given in Figure 1, the probabilities of "F_J" from v_1 to u_1, u_2, u_3 are $(1-d)(1-\alpha)C(u_1)$, $(1-d)(1-\alpha)C(u_2)$, $(1-d)(1-\alpha)C(u_3)$ respectively. The probability of "F_S" from v_1 to u_1 is $(1-d)\alpha$.

Note that we distinguish the action of changing or not changing the topic of interest only when the surfer is following a link. When jumping to a random page x, the surfer is always assumed to turn to a random topic (*jump-jump*, "J_J"). The reason behind this is that we consider jumping as a reset of the current browsing session or the start of a new browsing session. Therefore, it is intuitive to reset the preference distribution among topics as the static content distribution of the target page x. In the above example, the probabilities to "J_J" from v_1 to x_1, x_2, x_3 are $dC(x_1)$, $dC(x_2)$, $dC(x_3)$ respectively.

In summary, at each step of the random walk, the surfer may take any one out of the following three atomic actions: jumping to a random page and focusing on a random topic in the target page (action J_J); following a hyperlink and staying in the same topic (action F_S); following a hyperlink and jumping to any topic in that page (action F_J). Thus, the surfer's behavior can be modeled by a set of conditional probabilities.

$$\begin{aligned} P(F_S|v_k) &= (1-d)\alpha \\ P(F_J|v_k) &= (1-d)(1-\alpha) \\ P(J_J|v_k) &= d \end{aligned} \tag{1}$$

And the probability to arrive at topic i in target page u by the above actions can be described as

$$\begin{aligned} P(u_i|v_i, F_S) &= \frac{1}{O(v)} \\ P(u_i|v_k, F_J) &= \frac{1}{O(v)}C(v_i) \\ P(u_i|v_k, J_J) &= \frac{1}{N}C(u_i) \end{aligned} \tag{2}$$

The probabilistic model can be used to compute the probability that surfer is on page u for topic i, i.e., $A(u_i)$.

$$\begin{aligned} A(u_i) &= \sum_{v:v\to u} P((u_i|v_i, F_S)P(F_S|v_i)A(v_i)) + \\ &\quad \sum_{v:v\to u}\sum_{k\in T}(P(u_i|v_k, F_J)P(F_J|v_k)A(v_k)) + \\ &\quad \sum_{v\in G}\sum_{k\in T}(P(u_i|v_k, J_J)P(J_J|v_k)A(v_k)) \\ &= (1-d)\alpha \sum_{v:v\to u}\frac{1}{O(v)}A(v_i) + \\ &\quad (1-d)(1-\alpha)\sum_{v:v\to u}\frac{1}{O(v)}C(v_i)\sum_{k\in T}A(v_k) + \\ &\quad d\frac{1}{N}C(u_i)\sum_{v\in G}\sum_{k\in T}A(v_k) \end{aligned} \tag{3}$$

Let $A(v)$ denote the probability that a surfer at any time is browsing page v, with:

$$A(v) = \sum_{i\in T}A(v_i) \text{ and } \sum_{v\in G}A(v) = 1$$

Then Equation 3 can be simplified as:

$$\begin{aligned} A(u_i) &= (1-d)\sum_{v:v\to u}\frac{\alpha A(v_i) + (1-\alpha)C(v_i)A(v)}{O(v)} \\ &\quad + \frac{d}{N}C(u_i) \end{aligned} \tag{4}$$

After the propagation converges, each component $A(u_i)$ in the authority vector A_u:$[A(u_1), A(u_2), ..., A(u_T)]$ is the authority score of page u on topic i. $A(u)$ is the overall authority score. It can be proved that the sum of the topical authority scores $A(u_i)$ is identical to the one calculated by original PageRank algorithm. In other words, if the details of topic transitions are hidden, the model will reduce to the original PageRank.

From the analysis above, we can tell that authority distribution of a page does not only depend on its content, but also the topicality inherited from its ancestor pages.

3.3 Topical HITS

In the Topical HITS model, two importance vectors are associated with each page: the authority vector $A(u)$ and the hubness vector $H(u)$. Correspondingly, the random walk model in Topical HITS is a multi-surfer scheme based on the activity of two surfers: Surfer A and Surfer H. The authority vectors are generated by the activity of Surfer A while the hubness vectors are generated by the activity of Surfer H.

At each step, surfer A has two actions:

- F_S: follow an out-link and stay in the same topic with probability α
- F_J: follow an out-link and jump to any topic with probability $(1-\alpha)$

The probabilistic model can be formulated as

$$\begin{aligned} P(F_S|v_k) = \alpha \quad P(u_i|v_k, F_S) = \tfrac{1}{O(v)} \\ P(F_J|v_k) = (1-\alpha) \quad P(u_i|v_k, F_J) = \tfrac{C(v_i)}{O(v)} \end{aligned} \tag{5}$$

Surfer H has the opposite behavior with respect to surfer A:

- B_S: follow a back-link and stay in the same topic with probability α
- B_J: follow a back-link and jump to any topic with probability $(1-\alpha)$

And the probabilistic model can be formulated as

$$\begin{aligned} P(B_S|u_k) = \alpha \quad P(u_i|v_k, B_S) = \tfrac{1}{I(v)} \\ P(B_J|u_k) = (1-\alpha) \quad P(u_i|v_k, B_J) = \tfrac{C(v_i)}{I(v)} \end{aligned} \tag{6}$$

Moreover, the interaction between the surfers can be described based on the following mutual reinforcement relations:

$$\begin{aligned} H(v_i) &= \sum_{u:v\to u}P(v_i|u_i, B_S)P(B_S|u_i)A(u_i) + \\ &\quad \sum_{u:v\to u}\sum_{k\in T}P(v_i|u_i, B_J)P(B_J|u_i)A(u_k) \\ &= \sum_{u:v\to u}\frac{\alpha A(u_i) + (1-\alpha)C(u_i)A(u)}{I(u)} \end{aligned} \tag{7}$$

where $A(u) = \sum_{k\in T}A(u_k)$.

$$\begin{aligned} A(u_i) &= \sum_{v:v\to u}P(u_i|v_i, F_S)P(F_S|v_i)H(v_i) + \\ &\quad \sum_{v:v\to u}\sum_{k\in T}P(v_i|u_i, F_J)P(F_J|u_i)H(v_k) \\ &= \sum_{v:v\to u}\frac{\alpha H(v_i) + (1-\alpha)C(v_i)H(v)}{O(v)} \end{aligned} \tag{8}$$

where $H(v) = \sum_{k\in T}H(v_k)$.

If topical details are hidden, the model will reduce to the form of a normalized HITS, or a simplification of the SALSA model [12].

$$\begin{aligned} H(v) &= \sum_{u:v\to u}\frac{1}{I(u)} \times A(u) \\ A(u) &= \sum_{v:v\to u}\frac{1}{O(v)} \times H(v) \end{aligned} \tag{9}$$

The only difference between normalized HITS and original HITS is that a node's authority (hubness) will be distributed among its in-coming links (out-going links) to its parent (child) during propagation, while in original HITS, every in-coming link (out-going link) will get the entire authority (hubness) from the node.

3.4 Further discussion

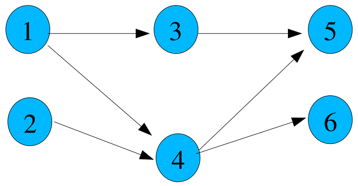
In the above model, α is the probability of the surfer to keep his interest when following an outgoing link, which is a tunable constant in experiment. However, a more reasonable consideration is that the decision about whether to keep a topic is usually dependent on the content of the current page. If this page is irrelevant to the topic of interest, the surfer is more likely to shift interest to another topic when entering a new page. In this case, α can be measured by the relevance of the topic to the the current content as $C(v_k)$, which is an variable instead of a constant. In the experimental section, we will check both options for α's setting.

Furthermore, hyperlinks can be divided into two types, intra-domain links and inter-domain links. The links that link two pages in the same domain are called intra-domain links, otherwise, they are called inter-domain links. According to the analysis in [10], the intra-domain links play less value than the inter-domain links when computing pages' reputation. We will investigate this issue by varying the relative weight of intra-domain links to inter-domain links (from 0 to 1) in our model.

3.5 A Simple Example

To understand how the topical model works, we take a initial look at a small web made up of six pages, as shown in graph 2(a). Each page is assigned a content vector across three pre-defined topics: Arts (A), Sports (S) and Business (B).

Interestingly, we assume page 5 doesn't contain any content, and as a result, a textual classifier generate a normalized distribution of (0.3,0.3,0.3) assuming such page's relevance to three topics are equal. Obviously such an assignment does not reflect the page's real topicality; however, our approach can help determine what the page is known for by using topical link analysis. As shown in Table 2(b), after running Topical PageRank, page 5 will obtain an authority vector 0.167, 0.065, 0.052, which means it is voted as the most authoritative page in the "Arts" category. Since there are two "Arts" pages (page 3 and page 4) cite page 5, it is reasonable to consider page 5 a good page about "Arts" instead of a page irrelevant to any topic. Similarly, page 6 is classified as a "sports" initially, but the "Arts" page linking (page 4) to it grants it some authority in "Arts". As a result, the topicality of page 6 will be a mixture of its static topicality and inherited topicality.



(a) A web made up of 6 nodes

node	content vector			authority vector			global auth.
	A	S	B	A	S	B	
1	0.2	0.7	0.1	0.019	0.065	0.009	0.093
2	0.2	0.4	0.4	0.019	0.037	0.037	0.093
3	0.9	0.1	0	0.102	0.031	0	0.133
4	0.7	0.3	0	0.119	0.081	0.013	0.213
5	0.3	0.3	0.3	0.167	0.065	0.052	0.283
6	0	1	0	0.0359	0.149	0	0.185

(b) Content and authority vectors

Figure 2: An example of topical model.

4. EXPERIMENTS AND RESULTS

To evaluate the behavior of our proposed topical link analysis algorithms, we compare the retrieval performance of well-known ranking algorithms versus the proposed topical link analysis algorithm. Comparisons are conducted among PageRank schemes and HITS schemes separately.

4.1 Data Collections

In their work on link spam identification, Wu and Davison [19] selected twenty queries from those used by previous researchers, ODP category names, and popular queries from Lycos and Google (shown in Table 1). For each query, they used search.yahoo.com to get the top 200 URLs; then for each URL, they retrieved the top 50 incoming links to this URL by querying Yahoo again. All pages referenced by these top 200 URLs are also downloaded. In this paper, we use this query-specific dataset to test various HITS-based ranking algorithms.

To evaluate PageRank-based ranking algorithms, we selected the TREC .GOV collection, which is an 1.25 million Web pages crawl of the .gov domain in the year of 2002. Among them, 1,053,372 are text/html files, which were used in our experiments. We chose the 2003 topic distillation task to test these algorithms, which contains fifty queries.

4.2 Textual Classification

We selected twelve top level categories from ODP as the broad topics. And a well-known naive Bayes classifier, "Rainbow" [16], is used to generate a content vector across these twelve topics for each document/query. The classifier is trained on 19,000 pages from each of the twelve categories of the ODP hierarchy.

```
web browser
jennifer lopez
table tennis
web proxy
trim spa
california lottery
US open tennis
wine
IBM research center
healthcare
rental car
super bowl
art history
weather
translation online
hand games
picnic
image processing
teen health
aerospace defence
```

Table 1: Set of twenty queries used for collecting query-specific datasets.

4.3 Experiments with HITS Schemes

In this section, we compare our Topical HITS (T-HITS with *imp* re-weighting) to traditional HITS [11], Bharat and Henzinger's *imp* (IMP) [2] and CLEVER's ARC weighting (ARC) [6] on the 20 query-specific datasets. Since our approach adopted out-link/in-link normalization in the calculation of authority/hubness, we also apply this hyperlink weight normalization on the above three approaches, generating the corresponding variations: normalized HITS (N-HITS), normalized IMP (N-IMP) and normalized ARC (N-ARC). As a result, there are seven approaches involved in the comparison.

4.3.1 Evaluation

Since there is no standard evaluation benchmark for the 20 query-specific datasets, the relevance between query and search results had to be inspected manually.

In our evaluation system, the top ten search results generated by various ranking algorithms were mixed together. To evaluate the performance, we enlisted a total of 43 participants, to whom a randomly chosen query and a randomly selected set of ten results (of those generated for the given query) were shown. The subjects were asked to rate each result as quite relevant, relevant, not sure, not relevant, and totally irrelevant, which were internally assigned the scores of 2, 1, 0, -1, -2, respectively. A page is marked as good if its average score across participants is greater than 0.5.

In this way, we can calculate the overall average precision (P@10) for each approach's top 10 results for 20 queries; in addition, the overall average score (S@10) is calculated to further explore the quality of retrieval since precision cannot distinguish top pages from merely good ones. We used these two metrics to compare the performance of various ranking approaches introduced above.

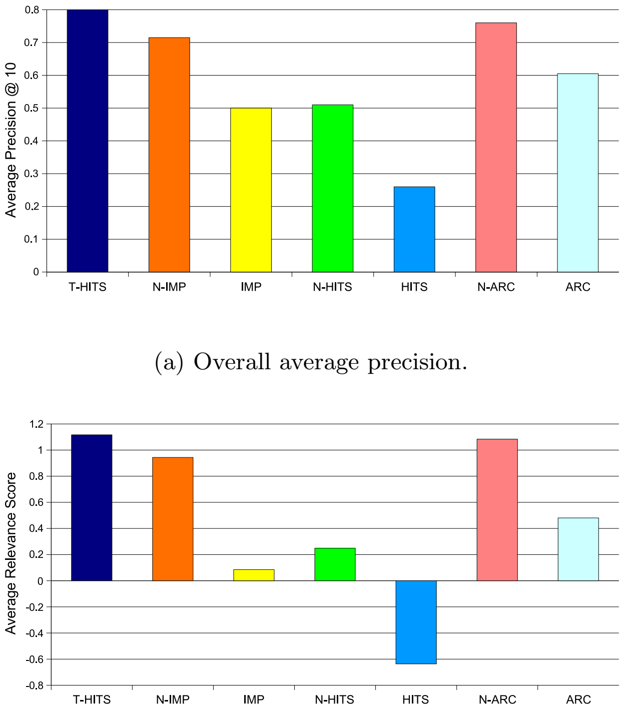
4.3.2 Results

Performance comparisons using precision and score are shown in Figures 3(a) and 3(b), respectively. With a precision of 80% and an average score of 1.12, our approach outperforms the other six. The second best is N-ARC, which gets 76% precision, with an average score of 1.08. N-IMP is ranked third with a precision of 71.5% and an average score of 0.94. Furthermore, we performed single-tailed t-tests to compare our T-HITS to N-ARC and N-IMP separately to study whether these improvements are statistically significant. The tests indicate that the improvement of T-HITS over N-IMP is significant (p-value=0.0005) while over N-ARC is not significant (p-value=0.26).

We noted that out-link/in-link normalization played a key role in boosting retrieval performance. This is because HITS-based approaches are vulnerable to link spam and the TKC effect [12], which will push pages within a tightly-knit community to high rankings even though the pages are not relevant, or pertaining to just one aspect of the topic. Such effect may derail HITS-based approaches and prevent them from finding relevant authorities. Normalizing link weights can alleviate such effects, thus providing better results.

4.4 Experiments of PageRank Schemes

In the above experiment, we show that our model works well on query-specific datasets. In this experiment, we find out whether the topical model retains its superiority when applied to global datasets as well.



(a) Overall average precision.

(b) Overall average human judgment scores.

Figure 3: Performance comparison of query-specific schemes.

Three ranking algorithms, traditional PageRank (PR) [14], topic-sensitive PageRank (TSPR) [8] and intelligent surfer (IS) [17] are compared to our proposed Topical PageRank (T-PR).

4.4.1 Evaluation

TREC provides relevance judgments for performance evaluation. We chose the topic distillation task in the web track of TREC 2003 to test these algorithms. The task contains fifty "topic distillation" queries with 10.32 relevant documents on average. We took P@10, Mean Average Precision (MAP) and Rprec [1] as the evaluation metrics as they are widely used in TREC.

4.4.2 Combination of IR score and importance score

We calculated the IR score with the OKAPI BM2500 [18] weighting function. Each document was represented by its full text plus the anchor texts of its incoming links. The parameters were set the same as Cai et al. [4],i.e., k_1=4.2, k_3=1000, b=0.8. We then chose the top 2000 documents according to BM2500 scores. For each approach, we constructed two different rankings on the 2000 documents: the ranking based on the BM2500 scores and the ranking based on importance scores obtained from this ranking algorithm. As a result, each document was associated with two order. We combined the two order as follows:

$$\gamma * rank_{IR}(d) + (1 - \gamma) * rank_{importance}(d) \qquad (10)$$

We then selected the top results from the combined list as

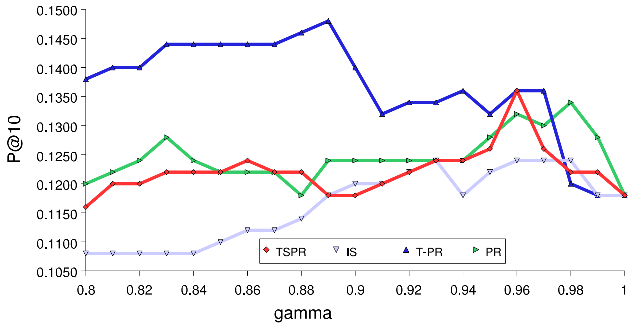


Figure 4: Combination of IR and importance scores.

the final outputs. Apparently, the parameter γ will impact performance; thus for each method, we tune γ to achieve the best performance.

4.4.3 Results

Figure 4 shows the precision@10 as γ is varied for the four approaches described above. As can be seen, T-PR curve is almost always above other curves in the graph, showing that Topical PageRank generally outperforms other approaches. T-PR combined with BM2500 gets the highest performance when γ is 0.89 with P@10 of 0.148. The TSPR combination gets the best result of 0.136 when γ is 0.96. Both PR and IS achieve their best performance when γ is 0.98, with P@10 of 0.134 and 0.124 respectively. All the curves converged to the baseline when γ is 1, which corresponds to the performance of BM2500 (P@10=0.118) weighting scheme. As can be seen from this figure, all ranking algorithms achieve higher performance than the baseline, which confirms that link analysis can improve performance on TREC data.

Figure 5 shows the overall performance comparison, where we selected the best result of P@10, MAP and Rprec for each approach. Topical PageRank exceeds other approaches on all of the metrics. An observation is that both TSPR and IS do not work well on TREC, as TSPR shows slight improvement over traditional PageRank, and IS performs even more poorly.

To determine whether these improvements are statistically significant, we performed several single-tailed t-tests. Table 2(a) shows the comparison of various approaches with the BM2500 baseline; only Topical PageRank performed significantly better than the baseline on all metrics at a 95% confidence level. For completeness, we also compared Topical PageRank to all other approaches. Even though the results, listed in Table 2(b), show that Topical PageRank only outperformed IS on the metrics of MAP and Rprec, we note that it is difficult to seek significant performance improvements given a topic distillation task where only a few relevant documents (10.32 on average) are associated with each query.

4.5 Parameter Tuning

In this section, we study how the parameters affect the performance of our proposed algorithms, where all parameters were tuned on TREC dataset. Figure 6 shows the variance of Topical PageRank's P@10 with different parameter settings.

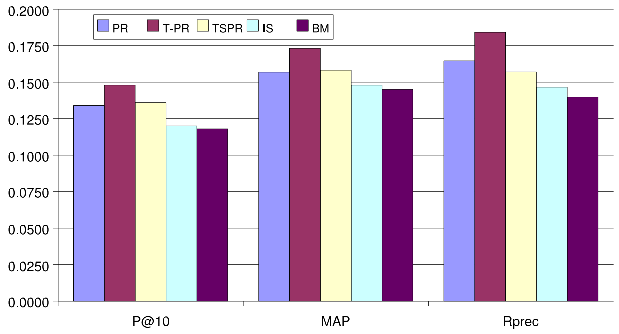


Figure 5: Comparison of overall performance.

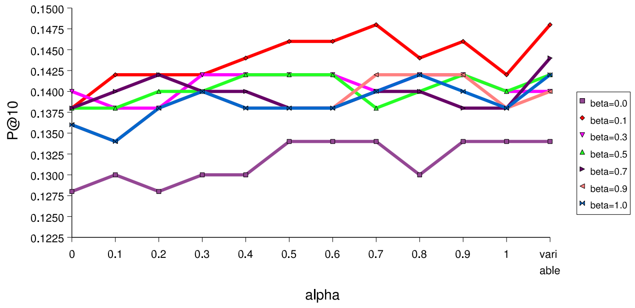


Figure 6: Precision @ 10 as alpha and beta are varied.

The parameter β is used to denote the relative weight of intra-domain link to inter-domain link. As shown in this graph, the curve with β=0.1 is almost always above other curves, which verifies the statement that the intra-domain links are less valuable and leveraging them to some degree can improve the performance. However, the performance decreased drastically if all intra-domain links are pruned (in the case where β=0); this is because over 80% hyperlinks in the web graph are intra-domain links, dropping them directly will destroy the global connectivity thus disrupting the authority propagation process.

In equations 3, 5, and 6, we used parameter α to represent the topical surfer's likelihood of staying in the same

Metric	T-PR	TSPR	IS	PR
P@10	**0.014**	0.086	0.130	0.073
MAP	**0.005**	0.036	0.067	0.068
Rprec	**0.012**	0.195	0.300	0.181

(a) Compared to BM2500

Metric	BM2500	TSPR	IS	PR
P@10	**0.014**	0.139	0.080	0.106
MAP	**0.005**	0.118	0.008	0.087
Rprec	**0.012**	0.220	0.018	0.147

(b) Compared to T-PR

Table 2: P-values in statistical tests.

topic when following a link, which can be either set as a constant within [0,1] or a variable measured by the topic's Naive Bayes score in the current page, as discussed before. An alternative explanation is that a page's topicality came from two factors, the contents of neighboring pages and its own contents, which are combined by α. From the graph, we can tell that such combination is necessary, since good result is always achieved in someplace between 0 and 1. Moreover, most curves get their best performance when α is set as the variable rather than a fixed constant.

In summary, our algorithm achieved the best result when α is set as the variable and β is 0.1, we used this setting in all of our experiments.

4.6 Discussion

An alternative interpretation of our work is that the per-node topical distributions with which we start and which is generated by link analysis is simply a representation in a reduced dimensionality space (equal to the number of topics). We start with vectors of length unity, but scale them based on link analysis authority calculations, and further adjust them via topical flows.

To examine this interpretation, we considered the performance of a simplified system—one that calculated, for a page u, a Static Topical PageRank score of ST-PR(u) = PR(u)C_u and a Static Topical authority score of ST-HITS(u) = N-HITS(u) C_u. We tested this formulation on the TREC dataset, and found that ST-PR achieved P@10 of 13.6%, which while not as good as our Topical PageRank score of 14.8%, still exceeded or matched all prior techniques. Similarly, ST-HITS achieved P@10 of 76.5% compared to our Topical HITS precision of 80%.

Therefore, we draw the conclusion that the topic distribution is an effective dimensionality reduction technique for content, but also that the flow of topicality in our topical link analysis techniques is necessary for top performance.

5. CONCLUSION AND FUTURE WORK

We have proposed a topical random walk model that probabilistically combines page topic distribution and link structure. We demonstrated how to incorporate this topical model within both PageRank and HITS without affecting the overall authority (or hub) score, and still provide a distribution of the authority across topics. Experiments on multiple data sets indicate that our algorithm outperforms a number of existing ranking approaches (both HITS-based and PageRank-based).

In the future, we expect to further study the effects of link weights (as in [9, 5, 4]). This is to include models of which links are more likely to be followed, or are of more value, or to assign a topic distribution to the links as well. We would also like to consider different kinds of topic distributions, e.g., fine-grained distributions (such as terms), coarse distributions (including binary classification such as spam/not-spam, or business versus educational), abstract distributions (like those formed as a result of dimension reduction).

Acknowledgments

We thank Baoning Wu for providing the twenty query-specific datasets. This material is based upon work supported by the National Science Foundation under Grant No. 0328825.

6. REFERENCES

[1] R. Baeza-Yates and B. Ribeiro-Neto. Modern information retrieval. In *Addison-Wesley Longman Publishing Co., Inc.*, Boston, MA, 1999.

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

[3] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. In *Proc. of the 7th Int'l World Wide Web Conf.*, pages 107–117, Brisbane, Australia, Apr. 1998.

[4] D. Cai, X. He, J.-R. Wen, and W.-Y. Ma. Block-level link analysis. In *Proceedings of the 27th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, July 2004.

[5] S. Chakrabarti, B. E. Dom, D. Gibson, J. M. Kleinberg, S. R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Mining the Web's link structure. *IEEE Computer*, pages 60–67, Aug. 1999.

[6] S. Chakrabarti, B. E. Dom, P. Raghavan, S. Rajagopalan, D. Gibson, and J. M. Kleinberg. Automatic resource compilation by analyzing hyperlink structure and associated text. In *Proc. of the 7th Int'l World Wide Web Conf.*, pages 65–74, Brisbane, Australia, Apr. 1998.

[7] Google, Inc. Google information for webmasters. Retrieved 9 November 2005 from the Google Website: http://www.google.com/webmasters/4.html, 2005.

[8] T. H. Haveliwala. Topic-sensitive PageRank. In *Proceedings of the Eleventh International World Wide Web Conference*, Honolulu, Hawaii, May 2002.

[9] IBM Almaden Research Center. The CLEVER Project. Home page: http://www.almaden.ibm.com/cs/k53/clever.html, 2000.

[10] K. M. Jiang, G. R. Xue, H. J. Zeng, X. Chen, W. Song, and W.-Y. Ma. Exploiting PageRank analysis at different block level. In *Proceedings of the 5th Conference on Web Information Systems Engineering*, 2004.

[11] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[12] R. Lempel and S. Moran. The stochastic approach for link-structure analysis (SALSA) and the TKC effect. In *Proc. of the 9th Int. WWW Conf.*, May 2000.

[13] Open Directory Project (ODP), 2006. http://www.dmoz.com/.

[14] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the Web. Unpublished draft, 1998.

[15] S. K. Pal and B. Narayan. A web surfer model incorporating topic continuity. *IEEE Transactions on Knowledge and Data Engineering*, 17:726–729, 2005.

[16] Rainbow: text classification tool. http://www.cs.umass.edu/~mccallum/bow/rainbow/.

[17] M. Richardson and P. Domingos. The intelligent surfer: Probabilistic combination of link and content information in PageRank. In *Advances in Neural Information Processing Systems 14*. MIT Press, 2002.

[18] S. E. Robertson. Overview of the OKAPI projects. *Journal of Documentation*, 53:3–7, 1997.

[19] B. Wu and B. D. Davison. Identifying link farm spam pages. In *Proc. of the 14th Int'l World Wide Web Conf.*, pages 820–829, Chiba, Japan, May 2005.

[20] Yahoo!, Inc. Yahoo! http://www.yahoo.com/, 2006.