# Separate and Inequal: Preserving Heterogeneity in Topical Authority Flows

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

## ABSTRACT

Web pages, like people, are often known by others in a variety of contexts. When those contexts are sufficiently distinct, a page's importance may be better represented by multiple domains of authority, rather than by one that indiscriminately mixes reputations. In this work we determine domains of authority by examining the contexts in which a page is cited. However, we find that it is not enough to determine separate domains of authority; our model additionally determines the local flow of authority based upon the relative similarity of the source and target authority domains. In this way, we differentiate both incoming and outgoing hyperlinks by topicality and importance rather than treating them indiscriminately. We find that this approach compares favorably to other topical ranking methods on two real-world datasets and produces an approximately 10% improvement in precision and quality of the top ten results over PageRank.

## Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

## General Terms

Algorithms, Performance

## Keywords

Web search engine, link analysis, reputation, PageRank

## 1. INTRODUCTION

Human recommendations are typically made within a particular context. For example, we say that person $P$ is a great plumber; it is rare to recommend someone without qualification (that is, without regard to purpose or topic). In addition, a person may also be known in multiple contexts: a successful scientist might also be an amateur musician.

Thus one might argue that reputation (or equally, importance or authority) is context-sensitive. This idea is equally applicable to the analysis of web authority. Traditional link analysis schemes treat hyperlinks indiscriminately and make the assumption that every target page is equally recommended. Under such methods, a page's importance is measured by the sum total of authority flowing through incoming links without considering from which communities that authority is derived. Similarly, in traditional link analysis methods, a page divides its authority equally among all targets without considering the context in which those links were created.

However, when a content creator places a hyperlink from one web page to another, often the link is meant to refer a particular aspect or topic of the target page instead of the page in general. Importantly, a single page often covers multiple topics, and so an indiscriminate analysis of an incoming link might mistakenly give importance to unintended topics.

In this paper we propose to model reputation within distinct domains. A page can be known in multiple areas, and thus we track a page's reputation separately within each domain. In addition, we consider to what reputation domain an outgoing link is headed, and propagate topical authority appropriately. To accomplish this, we break a node into a number of heterogeneous, domain-specific substructures and consider authority flow between those units (both within and between pages) instead of at the page-to-page level.

By modeling the structure of the web at this finer level of detail, we can separate authority flow on different topics and preserve such heterogeneity during authority propagation. As a result, a page's importance is able to be represented within multiple domains of authority, rather than by one that indiscriminately mixes reputations. Furthermore, by considering the relative similarity between the source and target domains, a page can propagate its topical authority among the targets in a more intelligent and appropriate way.

We evaluate our approach by conducting experiments on two real-world web datasets and involve studies on context representation as well as domain identification. We show that by introducing this fine-grained graph model, we can improve the retrieval quality of the PageRank algorithm.

The contributions of the paper are:

- We introduce a fine-grained link structure to better model authority propagation and the query-specific ranking.
- We examine in-page link context at multiple scales for domain classification.
- We propose in-page local transitions to preserve the heterogeneity in topical authority flows.
- We improve estimation of domain similarity by using a global context to describe domain-specific reputation.
- We compare our approach to a number of well-known ranking algorithms to show the superiority of our approach.

*SIGIR'08*, July 20–24, 2008, Singapore.

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 with link analysis and community identification. The proposed model is then detailed. The experimental framework and results will be presented in Sections 4 and 5 respectively. We conclude with a discussion and summary of contributions.

## 2. RELATED WORK

We begin with brief connections to related work, which can be put into four categories: web reputation representation, link analysis for community discovery, topicality in link analysis and web page partitioning.

### 2.1 Web page reputation

In this work we explicitly represent the various domains in which a page's reputation exists. Rafiei and Mendelzon [18, 17] also investigated the reputation of a page, and proposed a method to generate a description of a page's reputation by considering the contributions of all terms in predecessors that add authority to the page. In contrast, we consider distinct domain reputations within a set of topics, rather than a single cumulative reputation over all text. Our approach for generating reputation domains from link contexts is based on our prior work on identification of communities of parent pages [12, 13] but here we additionally consider extended anchortext and communities of child pages in order to maintain the heterogeneous flow of topical authority.

### 2.2 Link analysis for community discovery

Instead of classifying links into topical domains, one could partition links by the communities in which the link endpoints are placed. Processes for discovering communities on the web have been extensively studied. Kumar *et al.* [9] find expanded bipartite subgraphs to represent web communities. Similarly, Flake *et al.* [5] and Andersen and Lang [1] also detect communities purely based on link structure, while our approach takes page content into account when deciding communities. Moreover, these approaches utilize link analysis to detect web communities, while we use community information to assist with link analysis.

Roberts and Rosenthal [20] find clusters of web pages based on their outlink sets; in their model the authority value of a page is proportional to the number of clusters which tend to link to it rather than the number of pages which link to it. This is somewhat similar in spirit to our approach, except that we utilize page content when generating communities, and do not consider each community to be of equal value.

### 2.3 Topicality in link analysis

This paper is one of many that have proposed methods to integrate topical analysis with link analysis. Haveliwala's Topic-Sensitive PageRank (TSPR) [6] was the first algorithm to incorporate topical information into link analysis. In his model, multiple PageRank calculations are performed, one per category, each biased by jumping exclusively to pages in the corresponding category rather than to any web page. Rafiei and Mendelzon [18] apply a similar idea at the term level.

In addition to biasing the random jump on a per-term relevance basis, Richardson and Domingos's Intelligent Surfer (IS) [19] biases the selection among the outgoing links based on the relevance of the target to the term. Pal and and Narayan [16] similarly favor links leading to pages on the topic of interest.

In past work [11] we also incorporated topicality into web rankings in our Topical PageRank model. This approach calculates an authority score vector for each page to distinguish the contribution from different topics. A topical random surfer probabilistically combines page topical distribution and link structure.

All of the approaches above incorporate topics within the random surfer model. In contrast, the approach introduced in this paper retains the original random surfer model but applies it to a finer-grained link graph for better modeling of authority propagation though a page's reputation domains.

### 2.4 Partitioning web pages

Our proposed model partitions the set of outgoing links for each page (by domain). Others have similarly considered breaking a page into subpages. Chakrabarti *et al.* [3] propose segmenting a page into small pagelets containing contiguous subsets of links consistently on a particular topic. Cai *et al.* [2] use a vision-based algorithm to partition a page into blocks. In contrast, our proposed approach breaks the set of outlinks by the reputation domains to which they point, independent of the source page's structure.

## 3. METHODOLOGY

Web pages, like people, are often known by others in a variety of contexts. When those contexts are sufficiently distinct, a page's importance may be better represented by multiple domains of authority, rather than by one that indiscriminately mixes reputations. To achieve this, a page can be regarded as mapping into multiple "authority" units, in which each unit/subnode is used to accumulate authority flows from a particular domain. In this way, reputation flows from various domains will stay on their topics rather than diffusing into the entire page.

Similarly, a single page normally covers multiple topics; thus hyperlinks from different contexts within this page may point to different topics and carry different importance as well. Instead of giving each possible outlink the same amount of authority no matter whether the target's topic is relevant to the authority's context or not, it is advantageous to separate the targets into distinct domains and distribute authority among them based on relevance. Just as in the previous mechanism, we can separate the page into several domain-specific "hub" units, in which each hub includes outgoing links leading to targets in a particular domain.

In this way, we map the authority propagation from the web page level into a level consisting of domain-specific subnodes. Such a fine-grained model explores the hidden semantic structure within web pages between which a hyperlink really takes place, preserves the heterogeneity in topical reputation flow and provides a more effective method of reputation calculation. The rest of the section gives a detailed introduction of our model, Heterogeneous Topic Rank (denoted as HTR). For better understanding, we use a small web made up of four pages in Figure 1 as an example.

### 3.1 Identifying the various domains surrounding a page

As introduced above, the premise to identify the "authority" and "hub" units for a given page is to find out various distinct domains surrounding it. Web pages are often known by others or refer to others in a variety of contexts. To disambiguate various contexts, we classify them into separate categories. We predefine twelve broad categories (Table 1), chosen from the top level of the dmoz Open Directory Project (ODP) [14], and use a textual classifier to determine the category of each context. As shown in the second step in Figure 1, the classification process will assign a label (depicted by different colors in this example) to each hyperlink. In this way, given a page $u$, the hyperlink contexts pointing to and from it are

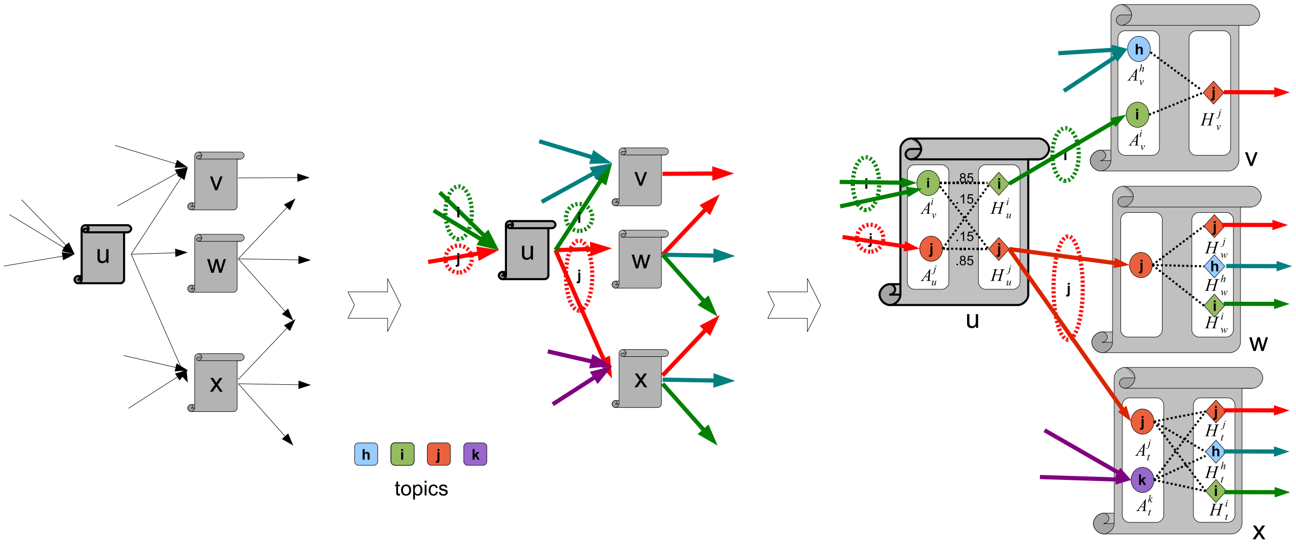

**Figure 1: The Heterogeneous Topic Ranking (HTR) process.**

separated into several domains (topics) based on their classification labels.

Besides the classification solution given above, we could also use clustering techniques to disambiguate contexts based on their textual similarity. Compared to classification, clustering would be more flexible. However, it is an expensive process, and in our case, would need to be applied to the set of parents for each document; even with parallelization and optimization, this will always be expensive for large datasets.

There are a number of options for representing a hyperlink's context. In this paper, we examine three variations:

- **Anchortext:** anchortext of the hyperlink.
- **ExtendedAnchor:** anchortext of the hyperlink plus the surrounding texts.
- **Fulltext:** full contents of the document where the hyperlink appears.

## 3.2 Exploring the fine-grained web graph

### 3.2.1 Node decomposition

With the various topics/domains surrounding a given page identified, we can easily partition the page from both directions, into domain-specific "authority" and "hub" units (denoted as $A$ and $H$ in what follows).

This process is demonstrated in the 3rd step of Figure 1. Node $u$ is linked from two relatively distinct domains, and domain $i$ represents 66% of the links and $j$, 33%. Node $u$ is split into two independent $A$ units, e.g., $A_u^i$ and $A_u^j$ (shown in the left rectangle within node $u$) with links from domain $i$ and domain $j$ are separately directed to $A_u^i$ and $A_u^j$. Symmetrically, we can decompose a page into several domain-specific $H$ units with regarding to various domains to which it points. As reflected in the example, since node $u$ links to two different domains $i$ and $j$, two independent $H$ units, e.g., $H_u^i$ and $H_u^j$ (shown in the right rectangle within node $u$) are introduced into node $u$ to separate outgoing links pointing to different domains. A similar process is applied to every node in the collection.

| Arts | Business | Computer | Games |
|---|---|---|---|
| Health | Home | Recreation | Reference |
| Science | Shopping | Society | Sports |

**Table 1: The twelve broad categories**

### 3.2.2 Representing $A$ and $H$ units

When node decomposition is complete, each page has been mapped into a bipartite graph consisting of two sets of vertices: units $A$ and $H$, with each $A$ (and $H$) associated with a particular domain. The decomposition processes of $A$ and $H$ units for a given page are symmetric, determined by various domains/topics pointing to or linked from this page respectively. To describe the domain referred by a particular $A(H)$, we use two levels of abstraction. Given an $A_u^i$ unit (shown in the 3rd step in Figure 1), we know that it is a unit within page $u$ referenced by a domain which contains several contexts on topic $i$. In "Term level" description, each context is represented in the form of a term vector, thus we can describe the domain by the average of all the belonging vectors, denoted as $t(A_u^i)$. Such an average vector (i.e., the centroid) integrates information from various members and renders a global view of this domain. Alternatively, since each context in the domain is labeled with the same category tag $i$, we can simply use the tag $i$ to denote the domain at "category level". The $H_u^j$ unit, in turn, can be represented using the textual average vector ("Term level") or classification label ("Category level") of all the $A$ units ($A_w^j$ and $A_t^j$ in our example) linked by it.

From the perspective of link analysis, the introduction of $A$ and $H$ units provides a fine-grained representation for hyperlink structure. In our model, each hyperlink is no longer a simple connection between two entire pages $u \rightarrow v$. Instead, it starts from some topical unit $H_u^i$ within the source page $u$, and points to the target topical unit $A_v^i$ rather than the complete page $v$.

### 3.2.3 Local flow within the page

In our model, we uncover the hidden internal structure of a page as a complete bipartite graph, with each $A$ unit linking to every $H$

unit within the same page. In this way, a particular $A$ can propagate its current authority to the outside via some intermediate "exits" ($H$ units). Rather than indiscriminately splitting authority among all the target domains no matter whether they are relevant to the domain of authority or not, a particular $A$ is more likely to grant its authority to an $H$ node on a similar topic, since this $H$ leads to targets in a relevant domain. As the example shows, the $A^i_u$ unit can either propagate its authority to the target domain $i$ ($A^i_v$) via $H^i_u$, or to target domain $j$ ($A^j_w$, $A^j_t$) via $H^j_u$, assuming that $i$ and $j$ are two unrelated domains. If we did not consider the relevance and equally split authority among the three outgoing links, 66% of authority on topic $i$ will divert into the irrelevant domain $j$. To avoid an unrelated domain getting undeserved high authority, we select $H$ units based on the relevance between the source and target domain. Suppose the relevance between $A^i_u$ and $H^i_u$ is 0.85, and 0.15 for $H^j_u$ (considering the occasional off-topic transition), reflected as the weight labeled on edges within the page. As a result, domain $i$ inherits 85% of the authority of $A^i_u$ while $j$ only gets a 15% share.

To measure the relevance $rel(A^i_u, H^j_u)$ between a pair of internal (A,H) units, e.g., $A^i_u$ and $H^j_u$, we calculate the similarity between their associated domains. As mentioned above, the domain can be described in two levels: the category level or the term level.

- **Category level:** In category level, if the involved $A$ and $H$ units are not labeled with the same classification tag, the transition probability is arbitrarily set to be a low value (0.15); the likelihood is non-zero to allow for the occasional off-topic shift.

$$rel(A^i_u, H^j_u) = \begin{cases} 0.85 & \text{if i == j} \\ 0.15 & \text{otherwise} \end{cases}$$

- **Term level:** In term level, both $A^i_u$ and $H^j_u$ are represented by the centroid of contexts within their domains, denoted as $t(A^i_u)$ and $t(H^j_u)$ respectively. The relevance can be measured by the cosine similarity of their vector space representations.

$$rel(A^i_u, H^j_u) = \frac{t(A^i_u)t(H^j_u)}{|t(A^i_u)||t(H^j_u|} \qquad (1)$$

In the last phase, we normalize the relevance measurement to represent the transition probability from $A^i_u$ to $H^j_u$, and use it to weight the corresponding edge in the page's bipartite graph.

$$nrel(A^i_u, H^j_u) = \frac{rel(A^i_u, H^j_u)}{\sum_{\tau:\tau\in D(H_u)} rel(A^i_u, H^\tau_u)} \qquad (2)$$

In the above equation, $D(H_u)$ denotes the collection of the category labels of $H$ units within page $u$.

## 3.3 Propagating authority

The next question comes as how to propagate authority among the newly constructed web structure. The solution is to imitate and alter the behavior of PageRank's "random surfer". Imagine a web surfer wanders on the web, who at each time is at some $A$ unit within a page $u$, e.g., $A^i_u$, and decides what is next to visit. With probability $d$, the surfer may jump to a randomly selected $A$ unit from the entire web with probability $1/N_A$, assuming that $N_A$ is the number of $A$ units in the whole collection. Otherwise, the surfer will pick one of the outgoing links, saying, the link towards $A^j_v$. However, notice that the outgoing links cannot be reached by the surfer directly, since there are several $H$ units blocking in the middle. The solution is to take a 2-step transition: the surfer first chooses an $H$ unit, and then follow one of its outlinks. The likelihood to choose a particular $H$ is determined by its relative relevance with the current $A$; the probability of selecting an outlink from $H$ is uniformly distributed. Let's take a look at the transition from $A^i_u$ to $A^j_v$. First of all, the surfer needs to hop from $A^i_u$ to $H^j_u$ within the page $u$, the relative likelihood is $nrel(A^i_u, H^j_u)$, as introduced in last section, which can be calculated in either "category level" or "term level". After arriving at $H^j_u$, the surfer can select uniformly one of its outlinks $H^j_u \rightarrow A^j_v$ with probability $1/O(H^j_u)$, where $O(H^j_u)$ is used to to denote the number of outlinks from $H^j_u$. In summary, the overall probability to transit to $A^j_v$ from $A^i_u$ is

$$(1-d)nrel(A^i_u, H^j_u)\frac{1}{O(H^j_u)}.$$

Since the authority score can be defined as the stationary probability of finding the random surfer at $A^i_v$, we can calculate $A^i_v$'s authority $HTR(A^i_v)$ as follows, with $D(A_u)$ denoting the collection of category labels of $A$ units within page $u$:

$$HTR(A^i_v) = (1-d)\sum_{u:u\rightarrow v}\frac{HTR(H^i_u)}{O(H^i_v)} + d\frac{1}{N_A} \qquad (3)$$

$$HTR(H^i_u) = \sum_{\tau:\tau\in D(A_u)} HTR(A^\tau_u)nrel(A^\tau_u, H^i_u) \qquad (4)$$

By replacing $HTR(H^i_u)$ in 3 with equation 4, the authority calculation can be finalized as

$$HTR(A^i_v) = d\frac{1}{N_A} + (1-d)\sum_{u:u\rightarrow v}\frac{\sum_{\tau:\tau\in D(A_u)} HTR(A^\tau_u)nrel(A^\tau_u, H^i_u)}{O(H^i_u)} \qquad (5)$$

When the authority propagation converges, every $A$ unit is assigned an authority score with respect to its associated domain. The next task is performed at query time; to be able to rank results for a particular query $q$, we need to calculate a query-specific importance score for the page. This can be achieved by merging the scores of $A$ units that belong to a page weighted by their affinity to this query. The composite score can be calculated as follows:

$$S_q(v) = \sum_{\tau:\tau\in D(A_v)} HTR(A^\tau_v) * r(q,\tau) \qquad (6)$$

where $HTR(A^\tau_v)$ is page $v$'s authority on domain $\tau$ and $r(q,\tau)$ represents $q$'s relevance to domain $\tau$. $r(q,\tau)$ is the $\tau^{th}$ component in query $q$'s probability distribution across the predefined categories, which can be generated by a textual classifier.

# 4. EXPERIMENTAL SETUP

In this section, we describe the datasets and experimental methods used to evaluate the performance. We will compare our approach versus well-known ranking algorithms, especially those combining text and link analysis. Experimental results will be presented in Section 5.

## 4.1 Datasets

To avoid a corpus bias, two different data collections were used in our experiments. One is the TREC GOV collection, which is a 1.25 million page crawl of the .gov domain in the year 2002. Among them, 1,053,372 are text/html files, which were used in our experiments. This corpus was used as the data collection of the TREC Web Track for a number of years. The second data set is a 2005 crawl from the Stanford WebBase [7, 4]. It contains 58M pages and approximately 900M links.

When conducting the experiments on the GOV corpus, we chose the 2003 topic distillation task to test these algorithms, which contains 50 queries. When doing experiments on the WebBase corpus, we selected 50 queries from a set of consisting of those frequently used by previous researchers, ODP category names, and popular queries from Lycos and Google (shown in Table 2).

## 4.2 Evaluation

Since there are no standard relevance judgments available for WebBase dataset, the relevance between query and search results has to be inspected manually. For each randomly assigned query, evaluators (members of our research lab) are required to determine the relevance for every URL result associated with this query (blind to the source ranking algorithm), using a five-value scale which were translated into the integers from 0 to 4. If the average score for this pair is more than 2.5, it is marked as relevant. The average fraction of relevant URLs within the top ten results across all queries is defined as Precision@10. To further explore the quality of retrieval, we also evaluated the ranking algorithms over the Normalized Discounted Cumulative Gain (NDCG) [8] metric. NDCG credits systems with high precision at top ranks by weighting relevant documents according to their rankings in the returned search results; this characteristic is crucial in web search.

For GOV data, TREC provides relevance assessments; there are 10.32 relevant documents per query on average for the topic distillation task of TREC 2003. In addition to P@10 and NDCG@10, we add Mean Average Precision (MAP) as evaluation metric since it is widely used in TREC and not restricted to the top 10 results.

## 4.3 Ranking methods compared

We compare our proposed approach HTR with five ranking algorithms: PageRank (PR) [15], Topical PageRank (TPR) [11], Topic-Sensitive PageRank (TSPR) [6], Intelligent Surfer (IS) [19] and CommunityRank (CR) [12]. PR is used as the baseline; we additionally chose TPR, TSPR, IS and CR because, similar to our model, they measure a page's reputation with respect to different aspects (topic or term) instead of a generic reputation.

As discussed previously, our model may have several options; the various resulting combinations are shown in Table 3. In the experimental section below, we study and compare their retrieval quality over multiple performance metrics.

| Method | Link Context | Domain Relevance |
|---|---|---|
| AC | Anchortext | Category |
| EC | ExtendedAnchor | Category |
| FC | Fulltext | Category |
| AT | Anchortext | Term |
| ET | ExtendedAnchor | Term |
| FT | Fulltext | Term |

**Table 3: Different HTR models.**

harry potter
music lyrics
online dictionary
olsen twins
weight watchers
playstation
new york fireworks
halloween costumes
st patricks day cards
the passion of christ
automobile warranty
herpes treatments
skateboarding
lord of the rings
angelina jolie
final fantasy
prom hairstyles
college football
george bush
britney spear
diamond bracelet
windshield wiper
jennifer lopez
moto racer
iraq war
four leaf clover
tattoos
fox news
paris hilton
taxes
hilary duff
star wars
janet jackson
musculoskeletal disorders
diabetes
nfl
pokemon
madonna
brad pitt
maps
poker
tsunami
games
jersey girl
golf clubs
pilates
seinfeld show
american idol
diets
poems

**Table 2: Set of fifty queries used for relevance evaluation in WebBase.**

For each query, we rank all documents using the combination of two different kinds of scores. One is the query-specific relevance score and the other is the authority score calculated by link analysis algorithms. The relevance score is calculated using the OKAPI BM2500 [21] weighting function, and the parameters are set the same as in Cai *et al.* [2]. We then select the top results from the combined list as the final outputs. The combination could be score-based, where a page's final score is a weighted summation of its authority score and relevance score; it could alternately be order-based, where ranking positions based on importance score and relevance score are combined together. In our implementation, we choose the order-based option; all ranking results presented in this paper are already combined with IR scores.

## 4.4 Textual classification

We use a well-known naive Bayes classifier, "Rainbow" [10], to decide the category for each hyperlink's context for the purpose of domain recognition as well as a given query's affinity to various topics. The classifier is trained on 19,000 pages from each of twelve categories of the ODP hierarchy. We apply it to both contexts and queries and get a topic distribution for each. We label the context by the dominant dimension of its topic distribution vector.

# 5. EXPERIMENTAL RESULTS

To evaluate the behavior of our proposed HTR model, we compare its retrieval performance versus well-known ranking algorithms. Results demonstrate that by preserving the heterogeneity in topical authority flows, we can produce more accurate search results.

## 5.1 Domain identification

As described in Section 3, each node in the web graph will be partitioned into several "authority" ($A$) and "hub" ($H$) units with respect to the different domains surrounding it. As a result, the 1.05 million nodes in GOV collection are mapped into 1.45 million $A$ units and 3.16 million $H$ units. The 54.7 million nodes in WebBase are mapped into 82.3 million $A$ units and 167.1 million $H$ units. (These numbers reflect the use of "ExtendedAnchor" context representation.)

We define *domain in-degree* as the number of domains pointing to a page, and *domain out-degree* as the number of domains pointed by a page. To illustrate, in Figure 2 we show the distribution of the domain in-degree and the domain out-degree when using extended anchortexts on GOV. (Distribution on WebBase exibits a similar pattern.) The distribution of domain out-degree is smoother than the domain in-degree, indicating that it is more common for a document to link to multiple topics than being known within many domains.

Many pages on the web are referenced by contexts from different domains. For example, in our GOV dataset, the

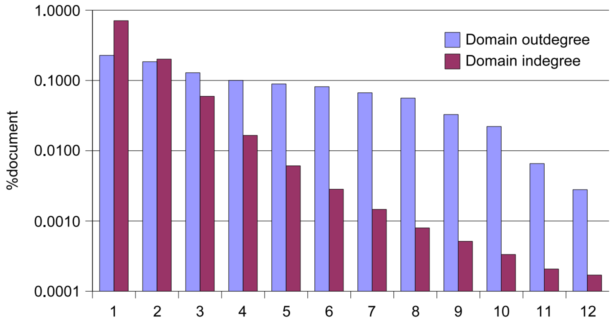

**Figure 2: Histogram of domain degrees for extended anchortexts on GOV.**

page http://www.tourism.wa.gov/ is the official site of Washington's state tourism. We classified its parent pages into either "Recreation" or "Business" domain. Another, http://oa.od.nih.gov/, is the home page of "NIH Office of Administration", for which some parents belongs to "Health", while others belongs to the domain of "Business". A broader example is http://dc.gov/, the governmental homepage of the District of Columbia. Information related to different topics can be found on this page, and correspondingly we found parent pages from various topics, such as "Recreation", "Sports", "Business", "Health" and "Computers", pointing to it.

## 5.2 Global context versus local context

When determining how to distribute the authority introduced through a particular incoming link among the targets, our model makes its choice based on the relevance of each source to each target. Instead of simply measuring the relevance based on the contexts provided by individual inlinks and outlinks, we choose to examine the relevance between domains. In this experiment, we investigate whether a global view of context (provided by a combination of all of the members of a link's domain) is more helpful than a local view of context (provided by a single link alone) in determining relevance.

We randomly sampled 966 linked pages from the GOV web document corpus. The relevance between source and target can be measured by the similarity between the link pair's local contexts (in "ExtendedAnchor" representation and with "term level" measurement). Unfortunately, the description provided by a single hyperlink is always short and not very informative, generating an extremely sparse term vector space. As a result, 604 out of 966 sample pairs end with zero relevance score, indicating that local context alone is not informative enough for relevance judgment. Global context provides a more detailed and comprehensive interpretation by synthesizing viewpoints from multiple members. When using global context, 323 out of 966 pairs have zero score. Compared to local context, the global representation reduces cases of zero by almost half or equivalently, increases cases of non-zero relevance scores by more than 75%. In this experiment, the average length of local context is 7.9 terms, versus an average of 187.1 terms for global context. In the remaining experiments, we use the global context to describe domain-specific reputation and estimate the inter-domain similarity.

## 5.3 Results on GOV dataset

The baseline performance (PageRank) on the three evaluation metrics introduced in Section 4.2 is shown at the top of Table 4.

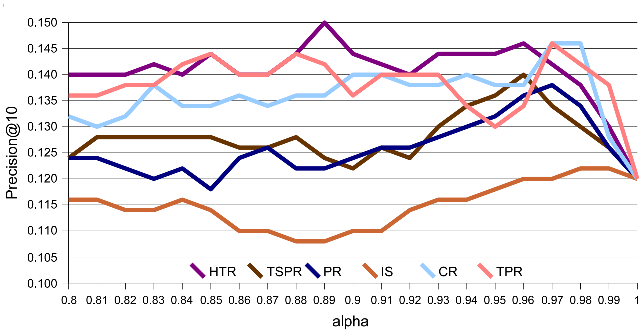

**Figure 3: Precision@10 performance on GOV as the combination of IR and importance scores are varied.**

Below the baseline we present performance of our model with different configurations. All but one result is better than the baseline. Thus we conclude that our model certainly has the potential to improve search quality. *HTR(EC)* gets the best performance on all measures. We also find that using "category level" to determine domain relevance outperforms or matches "term level" measurement in all cases.

Table 4 additionally shows (when using "category level") that using "ExtendedAnchor" to represent context exhibits better performance than using "Fulltext", while the performance of "Anchortext" lags behind. One possible reason is that the classifier performs poorly on short documents (i.e., anchortexts) since they are less informative and distinguishable, which may bring inaccuracy into domain identification and relevance measurement. Expanding the anchortext with surrounding information seems to improve performance to be even better than using "Fulltext".

Notably, when using "Fulltext" representation, only a single *H* unit is generated for each page because every outgoing link shares an identical contextual representation—the current page's full content. Since the outgoing targets cannot be discriminated in this case, the authority score is always equally divided among them, contradicting our intuition. Under this configuration, the HTR model degenerates into our previous model CommunityRank, and regardless of whether in "category level" or "term level" measurement, they exhibit the same performance as CommunityRank. Compared to "Fulltext", "ExtendedAnchor" provides a more efficient and specific way to represent hyperlink contexts.

After considering both quality and efficiency issues, we choose to use "ExtendedAnchor" to represent contexts in the following experiments.

In the following experiments, we compare the best performance of our model, with the other five rankers: PageRank, CommunityRank, Topical PageRank, Topic-Sensitive PageRank and Intelligent Surfer.

| Method | NDCG@10 | P@10 | MAP |
|---|---|---|---|
| **PR** | 0.218 | 0.138 | 0.153 |
| **HTR(AC)** | 0.232 | 0.138 | 0.167 |
| **HTR(AT)** | 0.219 | 0.132 | 0.165 |
| **HTR(EC)** | **0.243** | **0.150** | **0.174** |
| **HTR(ET)** | 0.222 | 0.144 | 0.156 |
| **HTR(FC)** | 0.240 | 0.146 | 0.168 |
| **HTR(FT)** | 0.240 | 0.146 | 0.168 |

**Table 4: Ranking performance of different HTR variations on GOV.**

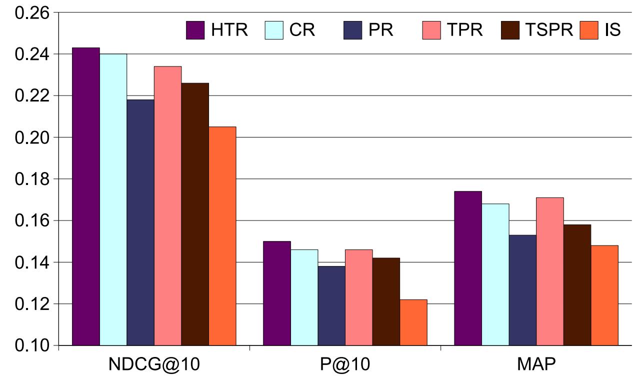

**Figure 4: Comparison of overall performance for GOV data.**

We first conduct a parameter study to investigate how different weights for importance and relevance scores will affect ranking systems' performance. Figure 3 shows the Precision@10 as $\alpha$ is varied for the four ranking approaches, where $\alpha$ is the weight of BM2500 score in the combination of text and authority. As can be seen, HTR curve is almost always equal to or above other curves in the graph, showing that our approach generally outperforms other approaches. All curves converged to the baseline when $\alpha$ is 1, which corresponds to the performance of BM2500. In GOV dataset, for each approach, we tune the combining parameter for the best P@10 and output its results with this optimal combination as final results. In contrast, for experiments on WebBase, we fix the weight of IR score as 0.8 to save the cost of manual evaluation across different values of $\alpha$.

Figure 4 shows the overall performance comparison. HTR outperforms other approaches on all metrics. An observation is that IS does not work well on TREC data, as it performs even more poorly than PageRank. To determine whether these improvements are statistically significant, we calculated several single-tailed t-tests to compare HTR with all other approaches. A t-test shows that HTR significantly exceeds both baseline and IS at a 90% confidence level.

## 5.4 Result on WebBase dataset

From experiments conducted on the TREC dataset, we drew the conclusion that using "ExtendedAnchor" to represent context provides appropriate descriptions and a significant cost savings compared to using full content. For WebBase, we only use "ExtendedAnchor" to represent context, but we still compare the two different options for relevance measurement: "category level" and "term level", as presented in Table 5. The baseline performance is listed in the top row.

Again, both the results presented in Table 5 are better than the baseline. *HTR(ET)* gets the best performance by outperforming the baseline by 10.8% on P@10 and 5.3% on NDCG@10. In contrast to the results shown in GOV dataset, "Term-Relevance" option outperformed "Category-Relevance" on WebBase.

| Method | NDCG@10 | P@10 |
|---|---|---|
| **PR** | 0.410 | 0.415 |
| **HTR(EC)** | 0.426 | 0.440 |
| **HTR(ET)** | **0.433** | **0.460** |

**Table 5: Ranking performance for different HTR approaches on WebBase.**

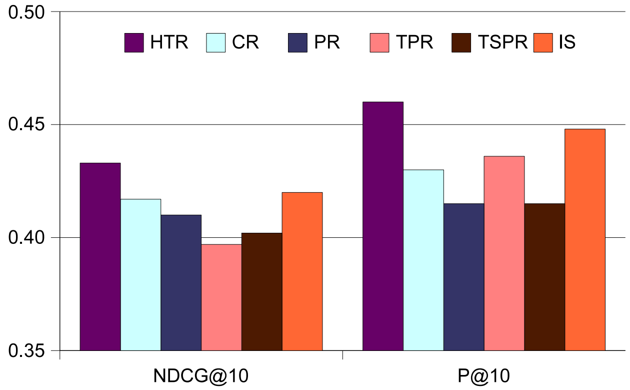

**Figure 5: Comparison of overall performance on WebBase data.**

Figure 5 shows the overall performance comparison. HTR outperforms the other approaches on both metrics. Again we performed t-tests to compare HTR with all the other approaches. A t-test shows that HTR significantly outperformed most approaches with a confidence level of at least 90% except for Intelligent Surfer. However, Intelligent Surfer is quite expensive, since it needs to be calculated for each dictionary term, while our model, like PageRank, only need to calculated once. In addition, different from HTR's consistent superiority on GOV and WebBase, Intelligent Surfer shows drastically different performance on the two datasets, from the worst to nearly the best.

# 6. DISCUSSION

From the above experiments, we find that "Term level" relevance measurement outperforms "Category level" measurement on WebBase, but not on GOV. Intuitively, queries in WebBase are broad and have lots of relevant documents, while queries in TREC are specific with only 10.32 relevant documents on average. As a result, there are different policies for "narrow" queries and "broad" queries. On one hand, we expand the similarity judgment from term-level to category-level for the purpose of including more potential candidates for the "narrow queries" on GOV; on the other hand, we focus on term-level to refine our search for "broad queries" used on WebBase. Some intermediate form, such as a finer-grained categorical representation, might be suitable for both scenarios, but is left for future work.

Since text vector space is sparse, it is no surprise that two reputation domains may not have significant overlap in text; on the other hand, even if two domains fall into the same broad category, the degree of relevance varies from case to case. A possible compromise is to combine the two relevance measurements so that we can decide whether the source and target are relevant based on "Category Level" results and further find out how relevant they are by examining the textual similarity at "Term level". In the future, we plan to explore a variety of ways to combine the two measurements.

Intelligent surfer exhibits quite poor performance on GOV dataset. A possible explanation is to note that intelligent surfer only wanders in a term-specific subgraph consisting of pages containing the particular term. Given a small dataset like GOV, it's hard to expect that such a graph will be well-connected and amenable to link analysis. Based on our statistics, the average density (links per node) of term-specific subgraphs in GOV (for terms in the 50 queries) is 3.11 versus 16.5 in WebBase.

# 7. CONCLUSION

In this paper we have proposed a novel ranking method that examines link contexts to divide the normal web graph into a finer-grained domain-based subnode graph. Our approach associates hyperlinks with particular reputation domains/topics and weights them with importance, so that multiple topical authority flows are propagated through the web graph heterogeneously. Experimental results on two real datasets indicate that this approach is consistently promising in improving search engines' ranking performance.

## Acknowledgments

This work was supported in part by a grant from the National Science Foundation under CAREER award IIS-0545875. We also thank TREC and Stanford University for access to their collections.

# 8. REFERENCES

[1] R. Andersen and K. J. Lang. Communities from seed sets. In *Proceedings of the 15th International World Wide Web Conference*, pages 223–232, Edinburgh, Scotland, May 2006.

[2] D. Cai, X. He, J.-R. Wen, and W.-Y. Ma. Block-level link analysis. In *Proc. of the 27th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval*, Sheffield, UK, July 2004.

[3] 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.

[4] J. Cho, H. Garcia-Molina, T. Haveliwala, W. Lam, A. Paepcke, S. Raghavan, and G. Wesley. Stanford WebBase components and applications. *ACM Trans. on Internet Technology*, 6(2):153–186, 2006.

[5] G. W. Flake, S. Lawrence, and C. L. Giles. Efficient identification of web communities. In *Proc. of the 6th ACM Int'l Conf. on Knowledge Discovery and Data Mining (KDD)*, pages 150–160, Boston, Aug. 2000.

[6] T. H. Haveliwala. Topic-sensitive PageRank. In *Proc. of the 11th Int'l World Wide Web Conf.*, pages 517–526. ACM Press, May 2002.

[7] J. Hirai, S. Raghavan, H. Garcia-Molina, and A. Paepcke. WebBase: A repository of web pages. *Computer Networks*, 33(1–6):277–293, May 2000. Proc. of the 9th Int'l World Wide Web Conf.

[8] K. Jarvelin and J. Kekalainen. IR evaluation methods for retrieving highly relevant documents. In *Proc. of the 23rd Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 41–48, July 2000.

[9] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Trawling the Web for emerging cyber-communities. *Computer Networks*, 31(11–16):1481–1493, 1999.

[10] A. K. McCallum. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. http://www.cs.cmu.edu/˜mccallum/bow, 1996.

[11] L. Nie, B. D. Davison, and X. Qi. Topical link analysis for web search. In *Proc. of the 29th Annual Int'l ACM SIGIR Conf. on Research & Development in Information Retrieval*, pages 91–98, Aug. 2006.

[12] L. Nie, B. D. Davison, and B. Wu. From whence does your authority come? Utilizing community relevance in ranking. In *Proceedings of the 22nd National Conference on Artificial Intelligence (AAAI)*, pages 1421–1426, July 2007.

[13] L. Nie, B. D. Davison, and B. Wu. Ranking by community relevance. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 873–874, July 2007.

[14] The dmoz Open Directory Project (ODP), 2008. `http://www.dmoz.com/`.

[15] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the Web. Technical report, Stanford University, 1998. Available from http://dbpubs.stanford.edu/pub/1999-66. Accessed 29 March 2008.

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

[17] D. Rafiei and A. O. Mendelzon. What do the neighbours think? Computing web page reputations. *IEEE Data Engineering Bulletin*, 23(3):9–16, Sept. 2000.

[18] D. Rafiei and A. O. Mendelzon. What is this page known for? Computing web page reputations. *Computer Networks*, 33(1-6):832–835, 2000. Proceedings of the 9th International World Wide Web Conference.

[19] 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.

[20] G. O. Roberts and J. S. Rosenthal. Downweighting tightly knit communities in world wide web rankings. *Advances and Applications in Statistics*, 3(3):199–216, Dec. 2003.

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