# Bridging Link and Query Intent to Enhance Web Search

Na Dai Xiaoguang Qi Brian D. Davison
Department of Computer Science & Engineering
Lehigh University
Bethlehem, PA, USA
{nad207,xiq204,davison}@cse.lehigh.edu

## ABSTRACT

Understanding query intent is essential to generating appropriate rankings for users. Existing methods have provided customized rankings to answer queries with different intent. While previous methods have shown improvement over their non-discriminating counterparts, the web authors' intent when creating a hyperlink is seldom taken into consideration. To mitigate this gap, we categorize hyperlinks into two types that are reasonably comparable to query intent, i.e., links describing the target page's identity and links describing the target page's content. We argue that emphasis on one type of link when ranking documents can benefit the retrieval for that type of query. We start by presenting a link intent classification approach based on the link context representations that captures evidence from anchors, target pages, and their associated links, and then introduce our enhanced retrieval model that incorporates link intent into the estimation of anchor text importance. Comparative experiments on two large scale web corpora demonstrate the efficacy of our approaches.

## Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

## General Terms

Algorithms, Performance

## Keywords

Link intent, Query intent, Kronecker product, Anchor text, Term weighting

## 1. INTRODUCTION

Search engine users issue queries with a variety of information needs, or intents. Some queries are targeted at finding particular web sites, while others are used to find generic information about certain topics. Some are issued to retrieve particular information that users have seen before, while some aim to explore new knowledge. Extensive work has been conducted to model, characterize, and understand such intents. One popular classification of query intent was proposed by Broder [3] in which queries are classified into three categories: informational, navigational, and transactional and are defined as follows (in Broder's original wording):

- Navigational. The immediate intent is to reach a particular site.
- Informational. The intent is to acquire some information assumed to be present on one or more web pages.
- Transactional. The intent is to perform some web-mediated activity.

A branch of research followed this classification scheme to predict user intent behind a query, and then utilize ranking models that can generate appropriate rankings for users. Such methods have shown improvement over their counterparts which do not discriminate query intent. However, most existing approaches have at least two limitations. First, they use hyperlinks and associated anchor text without differentiation. We argue that people create hyperlinks with different intents; therefore, such intent should be taken into consideration in these discriminative ranking approaches. Second, they only predict query intent based on the query itself. However, a query does not exist independently. It is connected with the retrieved documents, and with hyperlinks that are perhaps related to the query and the documents. Therefore, it is natural to collectively model the "intent" of different objects (queries, documents, and links) by utilizing their interconnections.

Given a web page, each of the hyperlinks that point to it is associated with an anchor text. As a preliminary step, we categorize hyperlinks into two types according to their intent: links that are created to describe the target page's identity (referred to as "navigational links"), and links created to describe the target page's content (referred to as "informational links"). For example, a link pointing to `http://www.pandora.com/` with the anchor text "Pandora" is considered a navigational link since the anchor text is the proper name of that particular internet music service. A link with the same anchor text pointing to the Wikipedia page about *Pandora*, the woman in Greek mythology, should be an informational link. Although the two classes of links seem to be mutually exclusive, there can be links that are a mixture of both. Still using the previous example where the target page is `http://www.pandora.com/`, if the anchor text is "Pandora, the internet radio", then it is partly a navigational link and partly an informational link. In our work, we will consider every link to be a soft combination of both, with the probability of being navigational and being informational sum to one.

Similarly to our classification of hyperlinks, we consider web documents as having the properties of attracting each type of link. Like hyperlinks, web pages can be a mixture of both. Note that both


*HT'11,* June 6–9, 2011, Eindhoven, The Netherlands.

the link types and web document types we propose correspond to the two primary query intent types: navigational and informational. Therefore, it is intuitive to emphasize different types of links and documents when generating rankings for different types of queries. In addition, such a correspondence makes it easier to model the interactions among the query, links and web documents.

Our work is conducted in two steps: first classify links into the two classes we proposed; then use the link intent classification results to generate better rankings. For link intent classification, we use a customized approach based on the Kronecker product of feature spaces that is capable of capturing the hidden interconnections between anchors and documents. We also include evidence from the link itself. Specifically, our evidence comes from three sources: (1) the anchor text string, from which popularity-based features are derived from statistics among associated target pages; (2) the target pages pointed to by the anchors; and (3) the hyperlinks connecting the anchors and targets. Given link intent classification result, we enhance retrieval models by incorporating it into the estimation of anchor text importance, to demonstrate its impact on web search. To the best of our knowledge, this is the first attempt to make an analogy of link intent with query intent[1], and incorporate this concept into retrieval models in a principled way. Experiments on two large scale web collections show our enhanced retrieval model achieves significant improvement over existing approaches. The contributions of this work include:

- We propose a classification scheme on link intent that benefits ranking performance, and investigate its rationality. We are not aware of any work by others on categorizing links by web authors' intent for retrieval improvements.
- We propose a feature-based model that exploits evidence from anchors, target documents, and hyperlinks to represent the context of a link. It effectively categorizes hyperlinks into our proposed scheme.
- We enhance anchor-based retrieval models by incorporating link intent classification results into the estimation of anchor text importance, and show its efficacy on ranking through thorough experiments.

The rest of this paper is organized as follows. We motivate this work and define the link intent classification problem in Section 2; introduce our problem approach for link intent classification in Section 3; present how we incorporate link intent into retrieval methods in Section 4; and report experimental results in Sections 5, 6 and 7. We review previous work in Section 8 and conclude our work in Section 9.

# 2. RATIONALITY OF LINK INTENT

In this section, we investigate the rationality of link intent on two aspects: (1) whether it is rational to categorize link intent into the classes of "navigational" and "informational"; and (2) why we expect incorporating this concept can affect ranking performance. We end by formalizing the link intent classification problem.

## 2.1 User Study on Link Intent

To verify the rationality of the proposed link intent taxonomy, we conducted a user survey on the motivation of a web user creating anchors that describe target pages. The survey investigated three main questions: (Q1) the purpose of a web author creating a link (with a short descriptive anchor text) to a *given* target page; (Q2) what type of target pages that a link (with a *given* anchor text) would point to; and (Q3) treating a *given* anchor text as a query for commercial search engines, what type of search results are expected. Our subjects are composed of 160 workers from Amazon Mechanical Turk[2]. Each subject is required to answer all three questions within one survey, and if legitimate through our manual check, she would get paid $0.3. Our objects are one anchor-URL pair per survey. We prepared the objects by (1) first randomly selecting around 800 anchor-URL pairs from WebBase [4, 11] (see Section 5.1 for details), and then (2) performing a manual check that filters out the unavailable links, resulting in 543 useful links. There are 502 valid survey answers returned finally.

**Table 1: Analysis of "Neither of above" answers.**

| Type | #. of resp. | Frac. |
|---|---|---|
| Informational links | 74 | 56.1% |
| Navigational links | 6 | 4.5% |
| Anchors not describing target pages | 38 | 28.8% |
| Others | 14 | 10.6% |

For Question 1, a majority of users chose a clear "informational" or "navigational" answer. However, it is interesting to see that 132 (26% of all valid answers) chose "Neither of above". We further investigated the text answers we collected for this question. We found that some users clicked the "Neither of above" choice by mistake; they knew the link should be informational or navigational, but clicked "Neither of above" unintentionally. Or, they are probably not aware that the link is informational or navigational; but from the text answer, we can tell that their understanding of those links fits our definition. Example answers of the above case include "a brief description of the given list", "yes it explains about the website", "The name of the website is x" (where x matches the anchor text). Considering the above factors, we found that more than 60% of the 132 "Neither of above" actually belong to informational or navigational links (56.1% and 4.5% respectively, as shown in Table 1). In another 28.8%, "Neither of above" is chosen is because the anchor text is irrelevant to the target page. For the remaining 10.6%, we could not get a clear reason why that choice is made. The statistics are shown in Table 1.

In summary, the results of this user survey verify our intuition that users can reasonably apply one of the two rationales we have provided for link intent in most cases.

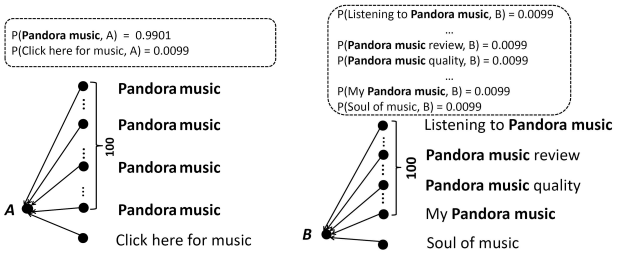
## 2.2 A Motivating Example on Ranking

The user study supports our intuition about link intent with real users' opinion. To explore the necessity of incorporating link intent into web search, we illustrate one example with two web pages in response to the query *pandora music*, as shown in Figure 1.

As mentioned in Section 1, *pandora music* is a navigational query, whose perfect answer is the home page of *pandora music* `http://www.pandora.com/`, i.e., page $A$ in Figure 1. Suppose both $A$ and $B$ associate with a series of anchor text respectively, our hope is to rank page $A$ higher than $B$. In the fictitious example in Figure 1(a), page $A$ associates with 100 in-coming links having the anchor text "pandora music", while $B$ in Figure 1(b) is pointed to by 100 inlinks with distinct anchor texts, all of which contain the substring "pandora music". Without considering page content and anchor text length, the contributions of anchor terms *"pandora music"* to page $A$ and $B$ are indistinguishable. However, we claim that page A should be emphasized when answering the navigational query "pandora music", and so we argue that differentiating link intents can help generate more accurate document representations for retrieval.

[1]Some preliminary results of this work is reported in [7].

[2]http://www.mturk.com/




(a) Page $A$ is the home page of *pandora music*.

(b) Page $B$ is a blog discussing music on Site *pandora music*.

**Figure 1: An example of two pages and their associated anchor text for answering *pandora music*.**

## 2.3 Problem Definition

We now formalize link intent classification problem. Four types of objects may hold specific intent within their contexts, i.e., queries (*Q*), documents (*D*), anchors (*A*) and links (*L*). In this work, an anchor refers to a unique string which may be used in multiple hyperlinks. Link intent is associated with its link context (*C(a,d)*), a function of the anchor and its pointed target page. Thus, we note that a link in our context is a unique anchor-document pair that maps to multiple hyperlinks on the web. We define the intent of the above objects as follows.

- A query is *navigational* if it aims to find *a particular site*; a query is *informational* if it aims to find *information about a certain topic*.
- A link is *navigational* if its anchor describes a target page's *identity*; a link is *informational* if its anchor describes a target page's *content*.
- An anchor is *navigational* if it mainly describes target pages' *identities*; an anchor is *informational* if it mainly describes target pages' *content*;
- A web document is *navigational* if it mainly attracts incoming links due to its *identity*; a web document is *informational* if it mainly attracts incoming links due to its *content*.

We formalize the link intent classification problem as follows.

***Problem Statement:*** *Given a link $l$ and its link context $C(a, d)$, determine whether its link intent is navigational or informational. The classification can be binary (hard) or probabilistic (soft).*

# 3. LINK INTENT CLASSIFICATION

In this section, we present our problem approach for link intent classification. It learns a feature based classification model based on evidence from link context profiles, and outputs a binomial distribution over every link, indicating the probability of belonging to a specific intent.[3]

## 3.1 Link Context Profiles

The observational data are naturally represented in a multi-dimensional format for each link context $C(a, d)$. We use the format of anchor×target page+link, in which the flatter form of anchor×target page (Kronecker product) is a one-way form (Figure 2) appending to the link dimension. Each dimension forms a space, in which features are extracted for profiling individually. We extract anchor, page (document) and link features to construct three types of profiles, represented as feature vectors $\overrightarrow{x_a}$, $\overrightarrow{x_d}$, and $\overrightarrow{x_l}$. These features are summarized in Table 2.

[3]We will present how we exploit such outputs to enhance retrieval models in Section 4.

**Table 2: Feature summary of link context profiles.**

| | |
|---|---|
| **Anchor Profiles (6)** | |
| KL(A) | KL divergence from anchor-link distr. to background. |
| En(A) | Entropy of anchor-link distr. |
| EnT(A) | Entropy of anchor term-link distr. |
| Diff(A) | Difference between top 2 popular items in anchor-link distr. |
| L(A) | Anchor length. |
| POS(A) | Fraction of noun., verb., etc. |
| **Page Profiles (8)** | |
| KL(P) | KL divergence from aggregated anchor-link distr. per page to background. |
| Diff(P) | Difference between top 2 popular items in aggregated anchor-link distr. per page. |
| En(P) | Entropy of aggregated anchor-link distr. per page. |
| EnT(P) | Entropy of aggregated anchor term-link distr. per page. |
| L(P) | URL length. |
| D(P) | URL depth. |
| **Link Profiles (5)** | |
| JC(T) | Jaccard coefficient between anchor and target page title. |
| JC(H) | Jaccard coefficient between anchor and target host. |
| IsApp(*) | Does anchor text appear in a specific field of target page, such as body, title, heading? |

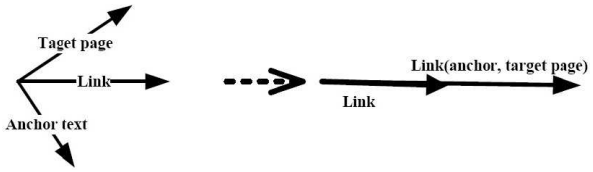



**Figure 2: The observational data have the relationship in which anchors and target pages can be treated as a 2-dimensional tensor, with its flat form appended to link dimension.**

*Anchor profiles* characterize anchor text from two aspects: (1) global anchor-link distribution, and (2) local anchor textual information. Given an anchor text, its anchor-link distribution gives evidence about its intent. For example, the links associated with a navigational anchor text (e.g., organization names) are more likely to point to pages in a navigational way. Lee et al. [16] proposed an effective feature to classify query intent from anchor-link distribution (Figure 3). Navigational pages tend to attract many links with the same anchor text that describes the page's identity. The anchor-link distribution is calculated as follows: (1) given an anchor, calculate how many times this anchor points to a given target page; (2) sort target pages according to anchor occurrence. Based on this distribution, we can extract much information characterizing the anchors, such as the distance of anchor occurrences between the target pages at rank one and rank two, its Kullback-Leibler Divergence to the anchor-link background distribution aggregated from all anchors, etc. Enlightened by Fujii [10], we calculate an entropy-based measure $i(a)$ to capture how skewed the distribution is. Let $D_a$ be the collection of target pages pointed by anchor $a$, the conditional entropy of $D_a$ is $H(D_a|a)$ is given by:

$$H(D_a|a) = -\sum_{d \in D_a} P(d|a) \log P(d|a)$$

To make $H(D_a|a)$ comparable among different $a$, we divided $H(D_a|a)$ by the logarithm of the size of $D_a$. Thus, $i(a)$ is normalized into the scale [0,1], defined as $i(a) = \frac{H(D_a|a)}{log|D_a|}$.

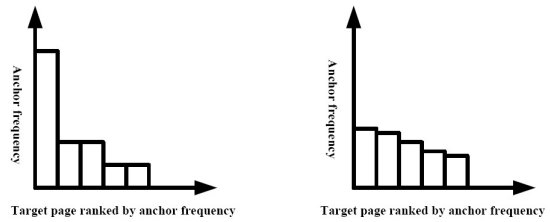



**Figure 3: An example of anchor link distribution: (left) navigational, (right) informational. From Figure 2 in [10].**

Since these factors are only based analysis of anchors, we added another group of factors similar to anchor-link distribution but based on anchor terms (referred to as "Anchor term-link distribution", see [10] for details). It combines the characteristics of other anchors that only share a portion of repeated anchor terms in a probabilistic way. Such features overcome the problem of anchor link distribution sparsity by generating a more confident distribution.

Textual features are also considered in profiling anchors. For example, we use the statistics based on the Part-Of-Speech (POS[4]) tags of anchor terms as features. Other features include anchor length, term frequency, and so on.

*Page profiles* characterize how likely a page is to attract links because of its identity. Given a target page, we aggregate the anchor-link distributions of all its in-linked anchors to profile the page intent. Define $A(d)$ as the collection of unique anchors pointing to $d$, the aggregated anchor link distribution on $d$, denoted as **Distr(d)**, is written as:

$$\textbf{Distr(d)} = \sum_{a \in A(d)} p(a|d) \cdot \textbf{Distr(a)}$$

where **Distr(a)** is the anchor link distribution of anchor $a$. Based on the distribution **Distr(d)**, the same group of features can be extracted to model page intent. Note that both anchor-level and term-level anchor link distribution are utilized in modeling page intent.

In addition to global features from **Distr(d)**, local features including URL length and URL depth, are also used in profiling. More features will refine the profile of target pages with respect to their intent.

*Link profiles* directly characterize the relationship between anchors and target pages. They are composed of features extracted from the direct comparison between anchors and target pages, such as the Jaccard coefficient between anchors and target titles, between anchor terms and target URL terms, etc.

## 3.2 Classification Model

Link context profiles are characterized on three different aspects, i.e., anchors, pages and links themselves. The format of the Kronecker product between anchor and page profiles enables modeling their hidden interactions. In this section, we present how we learn the link intent classification model based on such feature representation.

*Model Representation.* Let $A$ be the set of anchors, $D$ be the set of target pages, $L$ be the set of links connecting pairwise elements in $A$ and $D$. The intent on links is given by $S(L) \subseteq A \times D$, which means the link intent label $\widehat{s}_{ij}$ is given to the pair of $a_i$ and $d_j$ that has a link associated with them ($\forall a_i \in A$,$d_j \in D$). We define our classification model as:

$$\widehat{s}_{ij} = \sum_{i'=1}^{C_a} \sum_{j'=1}^{C_d} \widehat{w}_{i'j'} a_{i,i'} d_{j,j'} + \sum_{k'=1}^{C_l} \widehat{w'}_{k'} l_{ij,k'} \quad (1)$$

where $C_a$, $C_d$ and $C_l$ are the anchor, target page and link feature vectors respectively. $a_{i,i'}$ is the feature value of anchor $a_i$ at the $i'^{th}$ dimension, and $d_{j,j'}$ is the feature value of target page $d_j$ at the $j'^{th}$ dimension. $w_{i'j'}$ is the coefficient which represents the correlation between the anchor factor in dimension $i'$ and the target page factor in dimension $j'$. Note that it is independent with the instances of $a_i$ and $d_j$. Since the correlation between $a_i$ and $d_j$ is to fit the label on link $l_{ij}$ directly, the features on $l_{ij}$ can be reasonably combined with the anchor-page correlation linearly. We rewrite Equation 1 in matrix format:

$$\mathbf{S} = \mathbf{aWd^T} + \mathbf{w'l^T} \quad \text{or} \quad \mathbf{S} = \mathbf{dW^Ta^T} + \mathbf{w'l^T} \quad (2)$$

Considering the dynamics of the web graph, features from global anchor-link distributions may not be stable all the time. For example, when training and testing on web communities with many new sites which have few in-links to be able to form a confident anchor link distribution, the link profile is more important to determine link intent, and vice versa. When we further split $\mathbf{aWd^T}$ or $\mathbf{dW^Ta^T}$, we can also differentiate the importance of local features versus global features, which is sensitive to the anchor link distribution.

*Learning Process.* Our task is to learn the parameters $\theta = (\mathbf{W}, \mathbf{w'})$ by minimizing the prediction error $\sum_{ij}(s_{ij} - \widehat{s}_{ij})^2$. This can be done via maximizing log likelihood, denoted as $\arg\max_\theta P(S|A, D, L; \theta)$. Assuming labels on instances are independent, the likelihood is given by:

$$P(S|A, D, L; \theta) = \prod_{ij} p(s_{ij}|a_i, d_j, l_{ij}; \theta) = \prod_{ij} p(s_{ij}|\widehat{s}_{ij}) \quad (3)$$

We use a logistic regression model to estimate the likelihood $p(s_{ij}|\widehat{s}_{ij})$ via $p(s_{ij}|\widehat{s}_{ij}) = \frac{1}{1+\exp(1-s_{ij}\widehat{s}_{ij})}$. Fitted in a monotonic logarithm function, we maximize the logarithm likelihood in Equation 3, defined as $\arg\max_\theta \sum_{ij} \log p(s_{ij}|\widehat{s}_{ij})$. To avoid over-fitting, we add regularization objectives and rewrite our objective as

$$y = \arg\max_\theta \sum_{ij} \log p(s_{ij}|\widehat{s}_{ij}) - r_w \|\mathbf{W}\|_F^2 - r_{w'} \|\mathbf{w'}\|_F^2 \quad (4)$$

where $\|\cdot\|_F^2$ is the Frobenius norm, and $r_w$ and $r_{w'}$ are regularization parameters respectively.

Given the objective function, we calculate the differential of $y$ with respect to the model parameters $\theta$ as follows:

$$\frac{\partial y}{\partial w_{i'j'}} = \sum_{ij} \frac{\widehat{s}_{ij} \exp(1 - s_{ij}\widehat{s}_{ij}) a_{i,i'} d_{j,j'}}{1 + \exp(1 - s_{ij}\widehat{s}_{ij})} - r_w w_{i'j'}$$

$$\frac{\partial y}{\partial w'_{k'}} = \sum_{ij} \frac{\widehat{s}_{ij} \exp(1 - s_{ij}\widehat{s}_{ij}) l_{ij,k'}}{1 + \exp(1 - s_{ij}\widehat{s}_{ij})} - r_{w'} w'_{k'}$$

We choose to use gradient descent to estimate model parameters $\theta$ iteratively. Once we finish parameter learning, we predict the intent of a new link instance by Equation 1.

# 4. USING LINK INTENT FOR SEARCH

Web links reflect the intent of information providers, while queries represent the needs of information seekers. Therefore, connecting query intent with web link intent can help neutralize the inconsistency in interpreting information from different views in

[4]http://nlp.stanford.edu/software/tagger.shtml

web search. In this section, we present our enhanced anchor-based retrieval model by incorporating link intent. The general idea is that given the predicted link intent, we incorporate it into quantifying the importance of anchor text for document representation, and adapt it to query intent for retrieval. It actually adds another dimension of constraints on "intent" into the search process in which documents typically match the given queries lexically based on the evidence of term occurrence.

## 4.1 Modeling Anchor Text Importance

Exploiting anchor text to enrich document representations for retrieval has been widely studied in prior work. The underlying assumption is that anchor text is a short descriptive text that can provide complementary information for describing target pages. Earlier work ignored the distinguished importance of anchor text [6] or only utilized term frequency-based methods to quantify anchor text importance [24]. Craswell et al. [6] used the collection of anchor text as a surrogate document without differentiating anchor text importance. Westerveld et al. [24] modeled the importance of anchor terms by using $p(t|d)$, where $d$ is the surrogate documents only composed of in-linked anchor text. More recent works mitigated this deficiency by incorporating the knowledge from link structure. These works respectively are based on distinct assumptions. Fujii [10] extended anchor-based retrieval models by incorporating query intent inferred from anchor term link distribution, under the assumption that navigational queries benefit more from anchor-based models and informational queries benefit more from document-based models. Dou et al. [8] exploited site-level knowledge to de-emphasize the importance of anchor text associated with inlinks from the same site and cooperative sites. Metzler et al. [18] pointed out the problem of anchor text sparsity and enhanced anchor text representation by external anchor text and explored the effectiveness of diverse weighting strategies. Our work falls in a similar path, but differs from theirs in the sense that we differentiate links with different intent that are comparable with query intent and incorporate this into retrieval models.

Given a target page $d$, the importance of an anchor text $a$ can be modeled via:

$$f(a,d) \propto p(a,d) = p(a)p(d|a) = p(d)p(a|d) \tag{5}$$

where $p(a,d)$ is the probability that $a$ and $d$ have a certain relationship, such as having links associated with them. Note that $p(a,d)$ can be estimated in multiple ways. $p(a)$ and $p(d)$ are priors which reflect the probability that anchor $a$ and document $d$ appear on the web individually. The relationship between anchors and documents is estimated from $p(d|a)$ and $p(a|d)$. While $p(d|a)$ emphasizes the importance distribution from anchors to the associated documents, $p(a|d)$ implicitly indicates how to balance the contributions among different in-linked anchors to one target page.

By incorporating link intent, we add a new factor $i$ that represents link intent associated with anchor $a$ and target page $d$ in Equation 5, and so $p(a,d,i)$ can be modeled as:

$$p(a,d,i) = p(a)p(d|a)p(s(a,d) = i) \tag{6}$$

where $i \in \{$"info", "navi"$\}$. Thus, the original importance weight on each link is divided into two parts, with one proportional to $p(s(a,d) =$ "info") and the other proportional to $p(s(a,d) =$ "navi"). Such a weight distribution reflects the way that an anchor views its target page.

After computing the importance score of each individual anchor, we next combine all anchor text that points to the same page. For each target page, we collect all anchor text with their $p(a,d,$"info") scores. This collection indicates anchors' interpretation about page content combined with how likely they view the target page as informational. Such collective interpretation forms a type of view toward target pages. It is also applicable for the anchors with their $p(a,d,$"navi") scores. Hence, it is reasonable to separate such two collections into different fields of the target page. We call this **soft splitting** (denoted as **SS**). In contrast, we can divide anchor text deterministically according to $s(a,d)$. If one anchor is more navigational, it will be put into the page field entirely composed of navigational anchors. The same is applicable for informational anchors. Once determining which field an anchor should be in, its $p(a,d)$ score will be utilized as anchor importance with respect to the target page. We call this **hard splitting** (denoted as **HS**).

## 4.2 Intent-enhanced Retrieval Model

Each document is composed of three fields: (1) document content; (2) navigational anchor field; and (3) informational anchor field. We next combine multiple document fields into a unified retrieval model. We choose to use BM25F [21] since it can naturally incorporate the representations of multiple document fields into a single retrieval model. BM25F combines term frequencies in different fields linearly for BM25 score calculation. Suppose $w_f(i,j)$ is the weight of term $i$ for page $j$ in field $f$, it can be calculated by:

$$w_f(i,j) = \sum_{c \in f(j)} wt(c,j) \times tf(i,c) \tag{7}$$

where $wt(c,j)$ is the weight on component $c$ (unique anchor text in anchor fields) for the page $j$, and $tf(i,c)$ is the term frequency of $i$ in the component $c$. The aggregated term weights on $i$ is a linear combination of weights $i$ on all fields, which is given by:

$$w(i,j) = \sum_k \beta_k w_{f_k}(i,j) \tag{8}$$

where $f_k$ is the $k^{th}$ field of page $j$ and $\beta_k$ is a combination parameter, which controls the balance between term weights on each field used in BM25F ranking function ($\sum_k \beta_k = 1$). The document length is calculated by the same method.

The combination between two anchor fields and the document body field is trained automatically. The preference between two types of anchor fields can be either set by using query intent distribution or trained. We denote them as adaptive combination (denoted as **A**) and fixed combination (denoted as **F**) respectively. To achieve query intent distribution, we average the aggregated link intent associated with the in-coming links of the top $n$ search results returned by a ranking reference model (BM25 [22] in this work), where $n$ is 10 by default.

# 5. EXPERIMENTAL SETUP

## 5.1 Data Sets and Judgments

We conduct experiments on two large scale web corpora to avoid any bias from testbed data sets.

*ClueWeb.* TREC[5] provides standard relevance judgments on ClueWeb (Category B) for the evaluation of ranking algorithms. It contains 49.8M web pages and 940M hyperlinks approximately. We use the 50 queries (topics) within Ad hoc task of TREC 2009 Web Track for ranking evaluation.

To generate the anchor-document pairs used in link intent classification, we split the 50 queries into five folds sequentially by their IDs. We retrieve the top 2000 documents for each query by

[5]http://trec.nist.gov/

Okapi BM2500 [22] and randomly sample 200 inlinks pointing to these documents for queries in each fold as our examples. Note that there is no overlap between folds. In this way, we connect the selection of link examples to ranking characteristics. Each link example is labeled by at least one worker on Amazon Mechanical Turk, in the selection among "navigational link", "informational link", "both of them", and "none of them". To avoid links whose intents are uncertain (which may disturb classification accuracy), we only use the ones labeled as either "navigational link" or "informational link" (577 out of 1000 in total) as our examples for link intent classification.

*WebBase.* Our second data set is one 2005 web crawl from the Stanford WebBase [4, 11]. It contains approximately 58M web pages and 900M links. For ranking evaluation, 47 queries are selected from ODP category names, popular queries from commercial engines and those frequently used by previous researchers. we asked human editors (people in our lab) to assess the relevancy of a document to a given query, in selection among *not related*, *not relevant*, *borderline*, *relevant*, and *highly relevant*, which are translated into integer gains from 0 to 4.

To generate anchor-document pairs for link intent classification, we randomly sampled a few thousand links and manually labeled them as "navigational" or "informational". Like the case in *ClueWeb* where not all links have clear intents, we finally achieve 1281 anchor-documents pairs with definite intents and use them as our examples for link intent classification. We randomly split this data set into 10 folds.

## 5.2 Evaluation Metrics and Parameter Settings

We measure link intent classification performance on the metrics of $F_1$-measure and accuracy. $F_1$-measure is the harmonic mean of *precision* and *recall*. Here, *precision* is the percentage of truly positive examples in those classified as positive, while *recall* is the percentage of correctly classified positive examples out of all positive ones.[6] Accuracy is the percentage of correctly classified examples.

For *ClueWeb*, we use NEU methods [2] to measure ranking performance on statMAP and Precision at truncation level *k* (P@*k*), which is consistent with most prior work [5]. For *WebBase*, NDCG [13] is our metric. It penalizes irrelevant documents at top positions greater.

The combination parameter $\beta_k$ in BM25F is learned via hill climbing on metric statMAP (*ClueWeb*) and NDCG@10 (*WebBase*) respectively. For *ClueWeb*, ranking uses the same query splitting as link intent classification. For *WebBase*, ranking experiments are conducted based on two-fold cross validation, which is independent of the one used in the classification task.

## 5.3 Methods Compared

We compare a variety of baseline methods for both link intent classification and ranking tasks.

*Baseline methods for link intent classification.* We compare our link intent method with two groups of baselines, i.e., entity-based baselines and cluster-based baselines. The entity-based baselines exploited the following information:

- Anchor (Entity-A).
- Destination page (Entity-D).
- Anchor and destination page (Entity-AD).

[6]We will present link intent classification performance in Section 6, while varying the definition of positive examples.

The cluster-based baselines exploited the information of:

- Anchor (Cluster-A).
- Destination page (Cluster-D).

Entity-A uses the intent of anchors to predict the link intent. If an anchor has the entropy of anchor term link distribution less than 0.5, it is considered as a navigational query; we rank its target pages according to the $p(d|a)$, and generate a ranking list. The link to the top one page is considered as the only navigational link. Links to other target pages are informational.

Entity-D uses the intent of target pages to predict the link intent. If a target page has the entropy from aggregated anchor link distribution less than 0.5, it is considered as a navigational page, and we rank its associated anchors. The link associated with the anchor that has the highest $p(a|d)$ is considered to be navigational. All other links are considered to be informational.

Entity-AD averages the anchor term link distribution entropy of the anchor and the aggregated anchor term link distribution entropy of the target page to predict the link intent. If the average entropy is less than 0.5, the link is predicted as a navigational link; otherwise it is informational.

In Cluster-A, we cluster the links according to the anchor term link distribution entropy for anchors. For each anchor, we rank their target pages according to $p(d|a)$. The average of the anchor term link distribution entropies associated with target pages at a certain rank within the cluster is used to predict the link intent for all links at this rank.

In Cluster-D, we cluster the links according to the aggregated anchor term link distribution entropy of target pages. For each target page, we rank their anchors according to $p(a|d)$. The average anchor term link distribution entropy associated with anchors at a certain rank within the cluster is used to predict the link intent for all links at this rank.

*Baseline methods for ranking.* We compare our methods with three baselines:

- CDR: Ranking by Okapi BM2500 [22] based on the document body field.
- CQR: Ranking by [10] in which anchor and document-based models are combined by query intent.
- LinkProb: Dou et al. [8] utilizes link probability to weight anchor text (referred to as "LinkProb"), and then combines the weights into the BM25F model. Given a target page, the contribution of one unique anchor text line is proportional to the number of its associated incoming links.

# 6. EXPERIMENTAL RESULTS

We start by studying the performance of our link intent classification approach across different anchor-document feature representations. The performance on *ClueWeb* and *WebBase* are respectively reported based on 5-fold and 10-fold cross-validation. By picking up the best-performing classification model, we generate link intent labels on all anchor-document pairs on *ClueWeb* and *WebBase* respectively. We follow by performing comparative analysis on ranking evaluation.

*Performance on link intent classification.* We compare our link intent classification approach (denoted as **IntentC**) with baseline methods in Table 3. The ratio of navigational and informational links are 0.25 (*ClueWeb*) and 0.84 (*WebBase*). Preliminary results show the model renders the best performance when $r_w$ and

**Table 3: Link intent classification performance comparison on *ClueWeb* and *WebBase* data sets. *A*, *D* and *L* are feature vectors of anchor, document, and link respectively ($F_1$-meas.(navi) is the $F_1$-measure when navigational links are positive examples.).**

| | *ClueWeb* | | | *WebBase* | | |
|---|---|---|---|---|---|---|
| Methods | $F_1$-meas.(navi) | $F_1$-meas.(info) | Accu. | $F_1$-meas.(navi) | $F_1$-meas.(info) | Accu. |
| Entity-A | 0.162 | 0.827 | 0.714 | 0.662 | 0.782 | 0.723 |
| Entity-D | 0.186 | 0.814 | 0.698 | 0.669 | 0.776 | 0.722 |
| Entity-AD | 0.226 | 0.834 | 0.727 | 0.709 | 0.803 | 0.754 |
| Cluster-A | 0.148 | 0.834 | 0.722 | 0.742 | 0.728 | 0.734 |
| Cluster-D | 0.218 | 0.825 | 0.714 | 0.739 | 0.732 | 0.735 |
| IntentC(A+D+L) | 0.756 | 0.936 | 0.899 | 0.799 | 0.852 | 0.829 |
| IntentC(A×D×L) | 0.793 | 0.946 | 0.915 | 0.790 | 0.836 | 0.816 |
| IntentC(A×D+L) | 0.796 | 0.946 | 0.915 | 0.822 | 0.867 | 0.847 |

$r_{w'}$ are 0.1 and 0.01. We fix them and leave sensitivity analysis to future work.

Comparisons among baseline methods show that cluster-based methods do not provide an advantage under skewed class distribution. One possible reason is that the link instances with similar background have weaker capability of predicting the labels of target ones based on skewed class distribution. It especially influences the prediction of instances in the minority class, i.e., navigational links. Comparison within entity-based methods demonstrate that Entity-A and Entity-D perform the worst on accuracy. It is not surprising given both of them only utilize partial information of a link (anchor or document). By considering full information of a link, Entity-AD performs 2%-5% better than Entity-A and Entity-D on accuracy. For *WebBase*, cluster-based methods Cluster-A and Cluster-D incorporate the predictions on links with similar background, which outperform Entity-A and Entity-D by 1.5% and 1.8% respectively in terms of accuracy. Entity-AD outperforms Cluster-A and Cluster-D by around 1%-3%, indicating the combination of anchor term link distributions based on a link's two ends (anchor and document) contains stronger signals than the implicit influence among the links sharing similar background.

IntentC significantly outperforms baseline methods on all metrics for two data sets. Comparison on different feature representations shows the superiority of Intent(A×D+L) over both IntentC(A×D×L) and IntentC(A+D+L). One interpretation is that while Kronecker product of anchor and document feature spaces can capture the hidden interactions between anchors and documents when profiling link context, it also results in sparse feature vectors for many link instances, and so further exploiting a tensor product (three dimensions) that includes link feature space can make feature representations sparser and therefore remove some effective signals contained in link profiles.

In summary, IntentC(A×D+L) stably exceeds all baseline methods and its competitive variants, achieving reasonable performance on the link intent classification problem for both data sets. We next apply the model of IntentC(A×D+L) to generate link intent labels automatically, and investigate its impact on ranking in the rest of this section.

*Comparative performance on ranking.* Based on LinkProb, we incorporate link intent distributions while we vary how we separate anchor fields and how we determine the preference between two anchor fields. We denote our system variants as HS+F, HS+A, SS+F, and SS+A[7], and compare with baseline methods in Table 4. We conducted a single-tailed pairwise student t-test on ranking improvements. Significant differences over LinkProb are marked with †.

As expected, the baseline methods that exploit anchor text greatly outperforms those without. LinkProb and CQR have comparable ranking performance on both data sets. Our enhanced retrieval models (with four variants) outperform LinkProb and CQR on most metrics consistently, suggesting the effectiveness of link intent on improving anchor-based retrieval models. A closer look at our four variations exposes the following trends. First, soft splitting shows better overall ranking performance than hard splitting. One possible reason is that soft splitting smoothly enriches web pages by allowing links to play both roles with certain probabilities. Second, automatically adapting weights on navigational and informational anchor fields with query intent benefits the hard splitting approach, but hurts soft splitting. Our interpretation is that SS+A over-smooths the weights on anchor text in different fields of a target page, while HS+F fails to represent anchor text with respect to target pages naturally.

**Table 5: The top 5 search results generated by *LinkProb* and *SS+F* on query "fox news" (*WebBase*).**

| Query: fox news | |
|---|---|
| **LinkProb:** | |
| 1. | www.presidencia.gob.mx/vicentefox/ |
| 2. | www.presidencia.gob.mx/foxcontigo/ |
| 3. | www.presidencia.gob.mx/martadefox/ |
| 4. | www.aboriginalaustralia.com/catalog/product_info.php/products_id=227?oscsid=dc4f3b1303e8c41c000ff6a5fcdcfa3d |
| 5. | www.bridgeman.co.uk/about/collections.asp?type=&topic=956 |
| **SS+F:** | |
| 1. | **www.foxnews.com/** |
| 2. | www.presidencia.gob.mx/vicentefox/ |
| 3. | www.counterpunch.com/jacobs03202004.html |
| 4. | www.acsh.org/news/newid.316/news_detail.asp |
| 5. | www.counterpunch.com/leupp11292003.html |

## 7. DISCUSSION

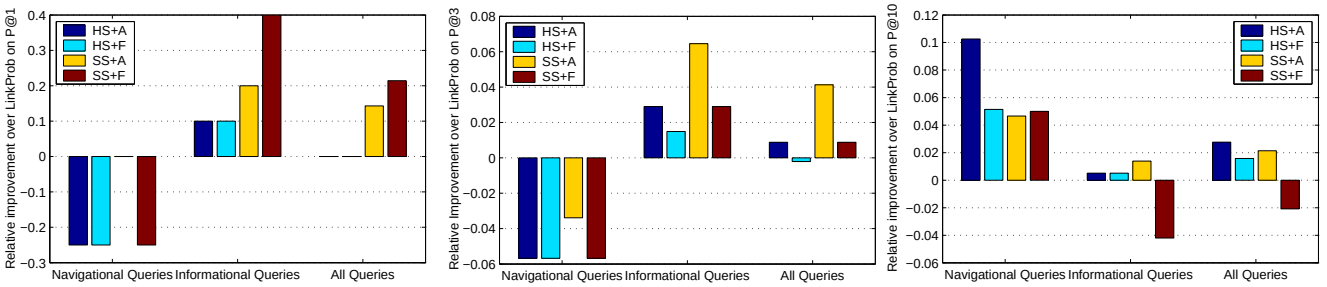
We showed that incorporating link intent improves anchor-based retrieval models. In this section, we will present a deeper analysis on the ranking improvements on different query types, e.g., query intent and query length. We also consider the effect of noise on link intent analysis by analyzing how ranking performance varies with the fraction of incorrectly classified link instances. We will also analyze feature effectiveness in link intent classification.

*Ranking improvement vs. query intent.* As discussed previously, link intent is similar to, and connected with, query intent in many ways. Therefore, we conjecture that the improvements made by incorporating link intent will be affected by the

[7]See Section 4 for the detailed explanation of each variant.

**Table 4: Ranking performance on *ClueWeb* (left) and *WebBase* (right) query sets. All methods are compared based on the parameter settings that achieve the best statMAP (*ClueWeb*) and NDCG@10 (*WebBase*) respectively. Performance with significant improvement (p-value<0.05) over LinkProb is marked as †.**

| | *ClueWeb* | | | | *WebBase* | | | |
|---|---|---|---|---|---|---|---|---|
| Methods | statMAP | P@1 | P@3 | P@10 | NDCG@1 | NDCG@3 | NDCG@5 | NDCG@10 |
| CDR | 0.175 | 0.204 | 0.224 | 0.304 | 0.148 | 0.161 | 0.167 | 0.171 |
| CQR | 0.176 | 0.286 | 0.306 | 0.375 | 0.398 | 0.355 | 0.349 | 0.369 |
| LinkProb | 0.175 | 0.285 | 0.306 | 0.377 | 0.407 | 0.351 | 0.356 | 0.371 |
| HS+F | 0.178 | 0.285 | 0.340† | 0.386 | 0.406 | 0.355 | 0.358 | 0.373 |
| HS+A | 0.176 | 0.285 | 0.340† | 0.390† | 0.408 | 0.358 | 0.359 | 0.373 |
| SS+F | 0.183† | 0.346† | 0.353† | 0.372 | 0.421† | 0.361 | 0.370† | 0.380 |
| SS+A | 0.179 | 0.326† | 0.306 | 0.387 | 0.410 | 0.376† | 0.354 | 0.377 |




(a) P@1 (b) P@3 (c) P@10

**Figure 4: The relative improvements over LinkProb on metric P@1, P@3 and P@10 for navigational queries, informational queries, and all queries on *ClueWeb*.**

intent of the query. Figure 4 shows the relative improvements on queries with different intents for *ClueWeb*[8]. Incorporating link intent brings greater improvements at top positions for informational queries than for navigational queries. This observation indicates that the link intent-enhanced retrieval model can make target page content representations more discriminative, and therefore help diversify search results. It is especially valuable for rankings of informational queries that are typically overwhelmed by a large number of relevant pages. For navigational queries, the enhanced retrieval models hurt ranking performance at very top positions but gradually improve ranking with the increase of truncation level (ranking position). This can be explained by the reason that an over-discriminative description about target pages makes search results at very top positions unstable. It especially hurts navigational queries which have only one best answer. We qualitatively verify our analysis through the example of the query "fox news" on *WebBase* in Table 5. Besides, it is worthwhile to point out that many techniques [6, 8, 10] have been proposed to improve rankings of navigational queries, which can mitigate this deficiency of our approach.

*Ranking improvement vs. query length.* Query length roughly reflects how narrow or clear users' information needs are. Table 6 shows the ranking improvement on queries with different lengths on both data sets. Intent-enhanced retrieval models bring more improvement for short queries, less improvement or even negative impact on long queries. This is not surprising given that short queries are more likely to be broad and ambiguous. Therefore they can benefit more from a more discriminative anchor representation. We consider this a useful property since most search engine queries are short [12].

[8]Topic 5, 15, 21, 23, 27, 31, 40, 41, 46 have navigational intent as revealed by manual inspection.

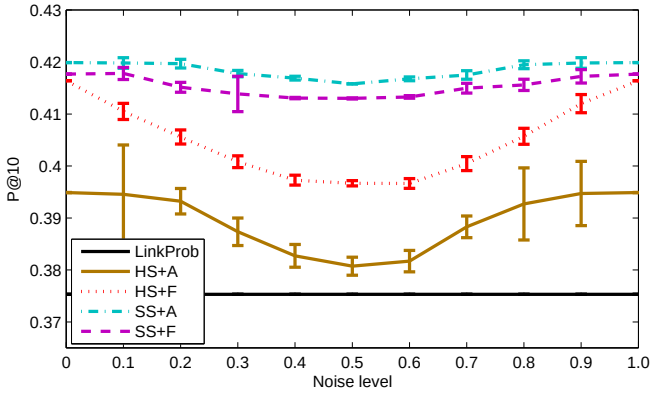



**Figure 5: Performance on metric P@10 on *ClueWeb* under different noise levels.**

*Analysis of noisy link intents.* To further analyze the impact of link intent on ranking improvements, we intentionally introduce noise into link intent by randomly sampling a fraction of links and reversing their link intent distribution. We run 30 times at each noise level. Figure 5 shows the average and standard deviation of ranking performance on P@10 under different noise levels on *ClueWeb*. The ranking performance of intent-enhanced retrieval

**Table 6: Ranking comparison on P@3 for *ClueWeb* (left) and NDCG@3 for *WebBase* (right). All methods are compared based on the parameter setting that achieves the best statMAP (*ClueWeb*) and NDCG@10 (*WebBase*). Performance with significant improvement (p-value<0.05) over LinkProb is marked as †.**

| | *ClueWeb* | | | *WebBase* | | |
|---|---|---|---|---|---|---|
| Methods | Length=1 | Length=2 | Length≥3 | Length=1 | Length=2 | Length≥3 |
| LinkProb | 0.156 | 0.313 | 0.458 | 0.384 | 0.340 | 0.321 |
| HS+F | 0.196(+25.0%)† | 0.392(+25.0%)† | 0.417(-9.09%) | 0.391(+1.78%) | 0.349(+2.50%) | 0.316(-1.67%) |
| HS+A | 0.196(+25.0%)† | 0.392(+25.0%)† | 0.416(-9.09%) | 0.398(+3.60%) | 0.346(+1.68%) | 0.316(-1.67%) |
| SS+F | 0.215(+37.5%)† | 0.352(+12.5%) | 0.479(+4.55%) | 0.413(+7.62%) | 0.405(+19.0%)† | 0.301(-6.36%) |
| SS+A | 0.156(+0.0%) | 0.274(-12.5%) | 0.479(+4.55%) | 0.402(+4.67%) | 0.351(+3.09%) | 0.333(+3.62%) |

models (with four variants) decrease with the increase of noisy link intent. Note that the trends are approximately symmetric with respect to noise level at 50% since retrieval models equally treat the two types of intent. Retrieval models are more robust and tolerant to noisy link intent when we (1) enhance anchor representation by softly splitting the contributions from navigational and informational anchor fields; and (2) adapt weights on navigational and informational anchor fields with query intent automatically.

*Feature analysis.* Link intent classification relies on features extracted from anchor, document and link profiles. We study feature effectiveness by examining the classification model generated by IntentC(A×D+L). For anchor and document profiles, features based on anchor/anchor term-link distribution (i.e., En(A), EnT(A) and Diff(A)) and the aggregated anchor/anchor term-link distribution per page (i.e., En(P) and Diff(P)) are the most effective. For the link profile, features based on comparison between anchor text and the target page title (i.e., JC(T) and IsApp(T)) are the most effective.

## 8. RELATED WORK

Connecting the properties of queries and anchor text has been well studied in previous work. Eiron and McCurley [9] investigated multiple properties of anchor text within a large intranet and showed its resemblance to real user queries in terms of term distribution and length. The work that exploited such resemblance developed into two directions. One direction utilizes anchor text properties to better answer a specific type of query. It is typically implemented through enriching document representations by anchor text. Craswell et al.'s work [6] on effectiveness of anchor text in site-finding tasks falls into this category. The other direction utilizes the properties of anchor text to better understand queries. Representative applications include query intent classification [16], query refinement [15] and query translation [17]. Our work differs from previous work by directly mapping query characteristics to anchor text (its associated links) and then considering effects on rankings through document representations. To achieve this, we draw from the techniques of query (intent) and link classification.

*Query intent classification* is one type of functional classification in which classifiers learn to determine the role of queries. As we mentioned earlier, there can be multiple classification schemes of query intent. Broder [3] suggested the three fundamental types of information need expressed in user queries. This classification scheme became popular in the query classification community. Following this path, Kang and Kim [14] proposed an approach to classify query intent into "topic-relevance" and "home-page finding" classes (i.e., "informational queries" and "navigational queries", respectively). Lee et al. [16] conducted a user study to demonstrate the viability of automatic query intent classification and proposed to identify query intent using "user-click behavior" and "anchor-link distribution".

*Link classification and link prediction* have been widely studied. Acar et al. [1] utilized a CANDECOMP/PARAFAC tensor decomposition to show the effectiveness of exploiting the natural 3-dimensional structure of temporal link data. Yu and Chu [25] utilized Gaussian process models for link prediction based on bipartite, direct and undirect graphs. Taskar et al. cast links as relational data and applied relational Markov network to model the joint distribution over the entire graph [23]. Qi et al. [20] analyzed the influence of link quality on web link-based algorithms, and proposed to classify links into two categories: those that confer authority and those that do not.

## 9. CONCLUSION AND FUTURE WORK

Query intent and link intent are implicitly connected. Revealing and exploiting such connections can benefit retrieval, and possibly many other tasks. In this paper, we proposed a method for automatic link intent classification based on evidence from anchors, target pages and the links themselves, and incorporated it into anchor-based retrieval models. We showed significant improvement upon the approaches that do not consider link intent.

This work can be extended in a variety of directions. Hyperlink classification could likely be improved by considering the source page and especially the hyperlink position within the source page. In terms of link classification taxonomy, we can also consider other taxonomies that connect queries with links, such as topics. While topical link analysis [19] has been well studied, it is still unclear the sensitivity of query and link taxonomy with respect to ranking quality. More generally, we search in a complex social network, in which people may create links for distinct purposes. How to associate users' information needs with the links connecting to appropriate resources is still an open issue. As a preliminary step, we connected query intent to link intent and showed such a connection is useful on web search. In the future, we plan to generalize the concept of link intent onto other entities and incorporate this to support social search.

## Acknowledgments

This work was supported in part by the National Science Foundation under awards IIS-0803605 and IIS-0545875, and an equipment grant from Sun Microsystems.

## 10. REFERENCES


[1] E. Acar, D. M. Dunlavy, and T. G. Kolda. Link prediction on evolving data using matrix and tensor factorizations. In *LDMTA'09: Proceeding of the ICDM'09 Workshop on Large Scale Data Mining Theory and Applications*. IEEE Computer Society Press, December 2009. In press.

[2] J. Allan, B. Carterette, B. Dachev, J. A. Aslam, V. Pavlu, and E. Kanoulas. Million query track 2007 overview. In *TREC*, 2007.

[3] A. Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, Fall 2002.

[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] C. L. Clarke, N. Craswell, and I. Soboroff. Overview of the trec 2009 web track. Technical report, NIST, no date.

[6] N. Craswell, D. Hawking, and S. Robertson. Effective site finding using link anchor information. In *Proc. of the 24th Annual ACM SIGIR International Conference on Research and Development in Information Retrieval*, pages 250–257, New Orleans, LA, Sept. 2001.

[7] N. Dai, X. Qi, and B. D. Davison. Enhancing web search with entity intent. In *Companion Proceedings of the 20th International World Wide Web Conference (WWW)*, pages 29–30, Mar. 2011.

[8] Z. Dou, R. Song, J.-Y. Nie, and J.-R. Wen. Using anchor texts with their hyperlink structure for web search. In *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 227–234. ACM, 2009.

[9] N. Eiron and K. S. McCurley. Analysis of anchor text for web search. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 459–460, New York, NY, USA, 2003. ACM.

[10] A. Fujii. Modeling anchor text and classifying queries to enhance web document retrieval. In *Proceeding of the 17th International Conference on World Wide Web*, pages 337–346, New York, NY, USA, 2008. ACM.

[11] 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 WWW Conf.

[12] B. J. Jansen, A. Spink, and T. Saracevic. Real life, real users, and real needs: A study and analysis of user queries on the web. *Information Processing and Management*, 36(2):207–227, 2000.

[13] K. Jarvelin and J. Kekalainen. IR evaluation methods for retrieving highly relevant documents. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 41–48, July 2000.

[14] I.-H. Kang and G. Kim. Query type classification for web document retrieval. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 64–71, New York, NY, 2003. ACM Press.

[15] R. Kraft and J. Zien. Mining anchor text for query refinement. In *Proceedings of the 13th International Conference on World Wide Web*, pages 666–674, New York, NY, USA, 2004. ACM.

[16] U. Lee, Z. Liu, and J. Cho. Automatic identification of user goals in web search. In *Proceedings of the 14th World Wide Web Conference (WWW)*, pages 391–400, New York, NY, 2005. ACM Press.

[17] W.-H. Lu, L.-F. Chien, and H.-J. Lee. Anchor text mining for translation of web queries: A transitive translation approach. *ACM Trans. Inf. Syst.*, 22(2):242–269, 2004.

[18] D. Metzler, J. Novak, H. Cui, and S. Reddy. Building enriched document representations using aggregated anchor text. In *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 219–226. ACM, 2009.

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

[20] X. Qi, L. Nie, and B. D. Davison. Measuring similarity to detect qualified links. In *Proceedings of the Third International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, pages 49–56, May 2007.

[21] S. Robertson, H. Zaragoza, and M. Taylor. Simple BM25 extension to multiple weighted fields. In *Proceedings of the 13th ACM Conference on Information and Knowledge Management*, pages 42–49, New York, NY, USA, 2004. ACM.

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

[23] B. Taskar, M. fai Wong, P. Abbeel, and D. Koller. Link prediction in relational data. In *Advances in Neural Information Processing Systems*, 2003.

[24] T. Westerveld, W. Kraaij, and D. Hiemstra. Retrieving web pages using content, links, urls and anchors. In *Proceedings of the 10th Text Retrieval Conference (TREC)*, pages 663–672. NIST, 2001.

[25] K. Yu and W. Chu. Gaussian process models for link analysis and transfer learning. In *Advances in Neural Information Processing Systems 20*, pages 1657–1664. MIT Press, 2008.