Building Contextual Anchor Text Representation Using Graph Regularization

Na Dai
Department of Computer Science and Engineering
Lehigh University
Bethlehem, PA 18015 USA

Abstract

Anchor texts are useful complementary description for target pages, widely applied to improve search relevance. The benefits come from the additional information introduced into document representation and the intelligent ways of estimating their relative importance. Previous work on anchor importance estimation treated anchor text independently without considering its context. As a result, the lack of constraints from such context fails to guarantee a stable anchor text representation. We propose an anchor graph regularization approach to incorporate constraints from such context into anchor text weighting process, casting the task into a convex quadratic optimization problem. The constraints draw from the estimation of anchor-anchor, anchor-page, and page-page similarity. Based on any estimators, our approach operates as a post process of refining the estimated anchor weights, making it a plug and play component in search infrastructure. Comparable experiments on standard data sets (TREC 2009 and 2010) demonstrate the efficacy of our approach.

Introduction

Human-generated web content interweaves through hyperlinks, referred to as *anchors*. When a web designer creates links pointing to other pages, she usually highlights a small portion of text on the current page, aiming to describe target page content or functionally link to target pages (e.g., “Click here”, “Last page”), and so facilitate visitors navigating to other information sources. Such highlighted text is referred to as *anchor text*.

Anchor text is usually succinct and descriptive, sharing similar properties with queries (Eiron and McCurley 2003) and interpreting target pages from the viewpoint of information describers (Dai, Qi, and Davison 2011). Anchor text has been widely applied to improve web search relevance (Fujii 2008; Dou et al. 2009; Metzler et al. 2009; Yi and Allan 2010; Dai and Davison 2010). The advantage is anchor text enables we estimate web page relevance based on an enhanced document representation. Here, the enhancement includes (1) the additional information introduced into target page content (Metzler et al. 2009; Yi and Allan 2010; Dai and Davison 2010), and (2) the intelligent ways of estimating anchor text importance (Dou et al. 2009). Our work makes efforts on the second aspect.

Most previous work on anchor importance estimation treated anchor text independently without considering its context. Anchor weights are simply accumulated based on their occurrence. Dou et al. (2009) and Metzler et al. (2009) both pointed this out and proposed to differentiate whether source and target pages are from the same site when quantifying anchor text importance. While these approaches verified the relationship between source and target sites is important, we argue that it is simply one type of useful relationship for improving anchor importance estimation. In particular, anchor weight estimation may benefit from three other types of broader relationship, i.e., that between (1) anchors pointing to the same page, (2) the anchor and its target page; and (3) similar anchors pointing to similar pages. Our essential goal is to stabilize the estimated anchor text weights from its context.

The anchor text collection of a target page comprises a description portfolio. Assuming better anchor weights positively correlate to ranking performance, our purpose is to maximize the quality of anchor text representation, and at the same time minimize the risk of anchor importance estimation. Here, the risk is important since it measures how confident our estimated anchor weights are. To quantify the risk, we necessarily draw from the interrelationship between anchors pointing to the same target page. There is an analog with the relationship between anchor text and target pages. Highly similar anchor text tends to capture target pages’ points, but may fail to provide complementary information.

Hyperlinks serve as recommendations to target pages (Brin and Page 1998). Assuming two similar pages have similar anchor text representation (Yi and Allan 2010), their similar anchor text tends to have consistent importance estimates. This idea actually has an analog in the area of information filtering and recommender systems, where similar items tend to attract similar rating distribution given a fixed group of users (Sarwar et al. 2001).

Based on the above concerns, we propose an anchor graph regularization approach, which quantifies these three types of relationship into constraints. To achieve this, we draw from the estimation of anchor-anchor, anchor-page, and page-page similarity from content-based and link-based viewpoints respectively. We incorporate the constraints into anchor text importance estimation and cast the task into a convex quadratic optimization problem. Based on any basic anchor importance estimators, our approach operates as a post process of refining the estimated anchor weights, making them more stable. Our contributions are two folds. First, we propose an anchor graph regularization approach to incorporate anchor text context into constraints for estimating anchor text importance. We are not aware of any previous work emphasized anchor text context for document representation. Second, we conduct comparable experiments on standard data sets (TREC 2009 and 2010 Web track (NIST 2010)), and demonstrate our approach achieves statistically significant ranking improvement over five representative baselines respectively.

The remainder of this paper is organized as follows. We first review previous work, followed by presenting how we quantify constraints to stabilize anchor importance estimation. We next introduce experimental setup and report our experimental results, followed by concluding our work with a discussion on future work.

Related Work

Using Anchor text to improve web search. Previous work has studied how to utilize anchor text for improving search relevance. (Craswell, Hawking, and Robertson 2001) is among the earliest, in which the authors demonstrated the effectiveness of anchor text on answering the information needs targeting at finding specific web sites. The following work on using anchor text to improve search gradually falls into three categories. One of them is to connect query intent with anchor text distribution on the web (Lee, Liu, and Cho 2005; Kraft and Zien 2004; Fujii 2008). Their observation is that the anchor text containing navigational query terms tends to have more skewed anchor-link distribution. It benefits web search in the way that we can use anchor text to customize ranking treatments for queries with different types of intent. The second category focuses on solving anchor text sparsity problem (Metzler et al. 2009; Yi and Allan 2010), i.e., only a few web pages has considerable amount of anchor text associated. The reason is that page in-coming links follow power law distribution (Amitay and Paris 2000). The effort within this category is to incorporate appropriate complementary anchor text to enrich existing anchor text representation. The third category focuses on intelligent ways of anchor text importance estimation. Dou et al. (2009)'s work that incorporated where source and target pages are from falls into this category. Our work is similar with theirs in that we both incorporate anchor text context into the importance estimation. However, the difference is that we focus on other types of broader relationship, and combine them into a unified optimization framework.

Modeling risks in information retrieval. Information retrieval (IR) community recently discussed risk-aware IR models (Wang and Zhu 2009; Wang 2009) for diversifying search results. Enlightened by mean-variance analysis of Modern Portfolio Theory (MPT), the risks of document relevance estimation capture the inter-document relationship, and are modeled to penalize the final relevance estimates. Empirical results demonstrated it achieves significant improvements over the baselines without considering relevance risk (Wang and Zhu 2009; Zhu et al. 2009). Modeling risk also succeeds in other IR tasks, including pseudo-feedback models and query expansion (Collins-Thompson 2008; 2009) and information fusion (Wang and Kankanhalli 2010). Similar with (Collins-Thompson 2008), we also formulate the risk as one of the optimization objectives. However, the fundamental difference is that we focus on anchor text representation, modeling the risk within anchor text importance estimation, while theirs focused on query representation, selecting the term set from pseudo-feedback documents for robust query expansion.

Utilizing similar pages in anchor text representation. Previous work has utilized similar web pages for their anchor text representation. Yi and Allan (2010) used the anchor text from semantically similar pages to enrich the anchor text representation of the current page. It is actually a smoothing process from anchor text of similar pages, mitigating anchor text sparsity problem and making the smoothed anchor representation more discriminative. Following the same spirit, we refer anchor text weights from similar web pages. The difference is we quantify anchor weight consistency into constraints, and solve this using graph regularization.

Anchor Graph Regularization

In this section, we present our problem approach for building a contextual anchor text representation. We first present the framework of anchor text importance estimation, and then introduce our parameter estimation.

Framework

Following previous work (Craswell, Hawking, and Robertson 2001; Westerveld, Kraaij, and Hiemstra 2001), we build up the surrogate document for each web page d from its associated anchor text collection $\mathcal{A}_d$. Here, $\mathcal{A}_d$ comprises all unique anchor text lines a, i.e., $\forall a \in \mathcal{A}_d$. Each a associates with one score $f_d(a)$, indicating its importance with respect to the target page d. $\mathbf{f}_d$ can be achieved from any basic estimator, e.g., (Fujii 2008; Dou et al. 2009; Metzler et al. 2009; Yi and Allan 2010). For simplicity, we define $\mathbf{p}_d$ as a probabilistic version of $\mathbf{f}_d$, i.e., $\mathbf{p}_d = \mathbf{f}_d / \|\mathbf{f}_d\|_1$, and formulate our task as

$$\begin{aligned} & \arg\min_{\hat{\mathbf{p}}_d} (1-\xi)\cdot\|\hat{\mathbf{p}}_d - \mathbf{p}_d^{(0)}\|_F^2 + \xi S(\mathcal{A}_d) \qquad (1) \\ s.t. \qquad & \|\hat{\mathbf{p}}_d\|_1 = 1 \end{aligned}$$

where $\hat{\mathbf{p}}_d$ is our estimates on anchor text importance, and $\mathbf{p}^{(0)}$ is the original estimates (i.e., outputs from one basic estimator), and $\|\cdot\|_F^2$ is the Frobenius norm. We conjecture that $\mathbf{p}^{(0)}$ serves as a reasonable expectation on anchor text importance. Therefore, we compute the deviation from $\mathbf{p}^{(0)}$ as part of the loss. Leveraging the loss from constraints $S(\mathcal{A}_d)$, we achieve the final anchor importance estimates by minimizing the sum of loss. $S(\mathcal{A}_d)$ is a set of constraints functioning on $\mathcal{A}_d$. In this paper, we focus on designing $S(\mathcal{A}_d)$ to incorporate anchor text context from (1) the relationship between different anchor text lines with a same target page, denoted

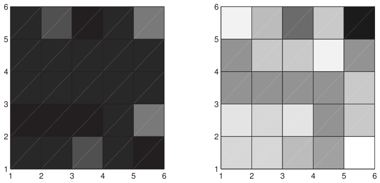


Figure 1: Anchor text weights associated with clueweb09-en0000-95-24509 in Clueweb09 data set before (left) and after (right) we apply constraint S_{aa}. The average weight on several most semantically similar anchor text lines for one grid. These grids are organized according to anchor-anchor similarity, so that more similar anchor text lines lay closer.

as S_{aa}; (2) the relationship between anchor text lines and the target page, denoted as S_{ad}; and (3) the relationship between similar anchor text lines associated with similar web pages, denoted as $S_{aa'}$. ξ is the trade-off controlling the emphasis on original estimates $\mathbf{p}_d^{(0)}$ versus bias to constraints $S(\mathcal{A}_d)$. When ξ is 0, the problem is trivial with the solution $\hat{\mathbf{p}}_d = \mathbf{p}_d^{(0)}$. Given $\hat{\mathbf{p}}_d$, we scale $\hat{\mathbf{p}}_d$ to achieve $\hat{\mathbf{f}}_d = \hat{\mathbf{p}}_d \cdot \|\mathbf{f}_d\|_1$. We now introduce how to formulate S_{aa}, S_{ad} and $S_{aa'}$.

Defining S_{aa}. The purpose of S_{aa} is to constrain the risks within anchor text importance estimation. Here, the risk comes from the situation that when most of the heavily weighted anchor text lines have large positive correlation, the anchor text representation may focus on a limited number of well representative content points. However, the risk that these points fall out of the target page coverage also increases, given that the anchor text representation is too focused, and may miss other important points that need to cover. To minimize such risk, we propose S_{aa} by using the covariance between different anchor text lines, which is enlightened by mean-variance analysis in Portfolio Theory. We define S_{aa} as:

$$S_{aa}: \qquad \arg\min_{\hat{\mathbf{p}}_d} \hat{\mathbf{p}}_d^T \mathbf{W}_d^{(aa)} \hat{\mathbf{p}}_d \tag{2}$$

where $\mathbf{W}_d^{(aa)}$ is the covariance matrix over anchor text lines in $\mathcal{A}_d$. We consider a special case here, i.e., the variance of individual anchor text line is 1, and interpret correlation by similarity measures[1]. Figure 1 shows the effect of S_{aa} on one document example. The observation is that anchor text weights are more diverse, especially for more similar anchor text lines.

Defining S_{ad} The purpose of S_{ad} is to leverage the coverage and balance of anchor text representation for a target page. Coverage measures how similar anchor text representation (weighted) are with the target page, while balance prevents any individual anchor text line from dominating the whole anchor text representation. The constraint on coverage is defined as

$$S_{ad}(cov): \qquad (\mathbf{w}_d^{(ad)})^T \hat{\mathbf{p}_d} \geq \lambda \tag{3}$$

where $\mathbf{w}_d^{(ad)}$ measures the similarities from anchor text line a to page d, and λ is a parameter serving as a lower bound on the weighted sum of similarity from each a to d. For balance, we define the constraint as:

$$S_{ad}(bal): \qquad \mathbf{w}_d^{(ad)}(a) \cdot \hat{\mathbf{p}}_d(a) \leq \varepsilon \quad \forall a \in \mathcal{A}_d \tag{4}$$

where ε is a upper bound on the similarity from individual a to d.

Defining $S_{aa'}$ The purpose of $S_{aa'}$ is to reference anchor text weights on similar pages for better determining the anchor importance on current pages. The main idea is to regularize on weighting consistency between paired anchor text lines on different pages. Given a target page d, we first search its top k similar pages (also known as the k nearest neighbors, denoted as $\mathcal{N}_d$, with $k = 20$), and then for $\forall d' \in \mathcal{N}_d$, we define the weighting constraints between d's anchor text line a and d''s anchor text line a'. The loss from inconsistent weighting is accumulated and minimized as part of the final optimization objectives. The constraint $S_{aa'}$ is defined as:

$$S_{aa'}: \arg\min \sum_{a \in \mathcal{A}_d} \sum_{d' \in \mathcal{N}_d} \sum_{a' \in \mathcal{A}_{d'}} c_{dd'}^{(aa')} (\hat{\mathbf{p}}_d(a) - \mathbf{p}_{d'}(a'))^2 \tag{5}$$

where $c_{dd'}^{(aa')}$ measures the similarity between a of d and a' of d', functioning on $\mathbf{W}^{(dd)}$, $\mathbf{w}_{d'}^{(ad)}$, and $\mathbf{w}_d^{(ad)}$ respectively. Here, $\mathbf{W}^{(dd)}$ is the similarity matrix between paired target pages. In this way, the weights on more similar anchor text lines describing similar pages are closer. We normalize $c_{dd'}^{(aa')}$ so that $\sum_{a \in \mathcal{A}_d} \sum_{d' \in \mathcal{N}_d} \sum_{a' \in \mathcal{A}_{d'}} c_{dd'}^{(aa')} = 1$ for a given page d. $c_{dd'}^{(aa')}$ is defined as:

$$c_{dd'}^{(aa')} = \frac{\mathbf{W}^{(dd)}(d,d') \mathbf{w}_{d'}^{(ad)}(a') \mathbf{w}_d^{(ad)}(a)}{\sum\limits_{a \in \mathcal{A}_d} \sum\limits_{d' \in \mathcal{N}_d} \sum\limits_{a' \in \mathcal{A}_{d'}} \mathbf{W}^{(dd)}(d,d') \mathbf{w}_{d'}^{(ad)}(a') \mathbf{w}_d^{(ad)}(a)} \tag{6}$$

Optimization. We presented how we formulate anchor text context as constraints and incorporate them into a unified optimization framework, we now discuss how to solve this optimization problem. We summarize our problem as follows.

$$\begin{aligned}
& \arg\min_{\hat{\mathbf{p}}_d} (1-\xi)(1-\mu) \|\hat{\mathbf{p}}_d - \mathbf{p}_d^{(0)}\|_F^2 \qquad (7) \\
& +\xi(1-\mu) \hat{\mathbf{p}}_d^T \mathbf{W}_d^{(aa)} \hat{\mathbf{p}}_d \\
& +\mu \sum_{a \in \mathcal{A}_d} \sum_{d' \in \mathcal{N}_d} \sum_{a' \in \mathcal{A}_{d'}} c_{dd'}^{(aa')} (\hat{\mathbf{p}}_d(a) - \mathbf{p}_{d'}(a'))^2 \\
s.t. \quad & \|\hat{\mathbf{p}}_d\|_1 = 1 \\
& (\mathbf{w}_d^{(ad)})^T \hat{\mathbf{p}}_d \geq \lambda \\
& \mathbf{w}_d^{(ad)}(a) \cdot \hat{\mathbf{p}}_d(a) \leq \varepsilon \quad \forall a \in \mathcal{A}_d
\end{aligned}$$

This is a standard quadratic optimization problem. $\mathbf{W}_d^{(aa)}$ is actually an anchor-anchor covariance matrix, which makes it a positive definite matrix. The diagonal matrix constructed by $\sum_{d' \in \mathcal{N}_d} \sum_{a' \in \mathcal{A}_{d'}} c_{dd'}^{(aa')}$ is also positive definite. Therefore, our optimization problem is guaranteed to be convex, which has a global minimizer if there exists some feasible solution (satisfying constraints). To guarantee all conditions can

[1] "Correlation" and "similarity" are interchangeable here.

be satisfied, we incorporate slack variables that measure violation of the conditions, and formulate them into existing objectives. We transform Equation 7 into:

$$
\begin{aligned}
& \arg\min_{\hat{\mathbf{p}}_d} (1-\xi)(1-\mu)\|\hat{\mathbf{p}}_d - \mathbf{p}_d^{(0)}\|_F^2 \\
& + \xi(1-\mu)\hat{\mathbf{p}}_d^T \mathbf{W}_d^{(aa)} \hat{\mathbf{p}}_d \\
& + \mu \sum_{a \in \mathcal{A}_d} \sum_{d' \in \mathcal{N}_d} \sum_{a' \in \mathcal{A}_{d'}} c_{dd'}^{(aa')} (\hat{\mathbf{p}}_d(a) - \mathbf{p}_{d'}(a'))^2 \\
& + \lambda_1 + \sum_{a \in \mathcal{A}_d} \varepsilon_1(a) \\
s.t. \quad & \|\hat{\mathbf{p}}_d\|_1 = 1 \\
& (\mathbf{w}_d^{(ad)})^T \hat{\mathbf{p}}_d \geq \lambda - \lambda_1 \\
& \mathbf{w}_d^{(ad)}(a) \cdot \hat{\mathbf{p}}_d(a) \leq \varepsilon + \varepsilon_1(a) \quad \forall a \in \mathcal{A}_d \\
& \lambda_1 \geq 0 \\
& \varepsilon_1(a) \geq 0 \quad \forall a \in \mathcal{A}_d
\end{aligned}
\qquad (8)
$$

The ways of solving convex quadratic optimization problems include interior point, active set (Murty 1988), augmented Lagrangian (Delbos and Gilbert 2003), conjugate gradient, and etc. Given that we face on the challenge of processing large-scale data, we choose to use interior point method implemented in OOQP-0.99.22 (Gertz and Wright 2003) with its default solver.

Parameter Estimation

We presented how we optimize the problem. We now introduce how to estimate parameters, i.e., anchor-anchor similarity ($\mathbf{W}^{(aa)}$), anchor-doc similarity ($\mathbf{w}^{(ad)}$) and doc-doc similarity ($\mathbf{W}^{(dd)}$). Two main steps are (1) representing anchor text lines/documents and (2) computing similarity measures.

We start by introducing how we represent anchor text lines and web pages. For anchor text line a,

- **Content-based representation**: select k' web pages that have the most in-coming links associated with a. Concatenate the content of these k' web pages to represent a, where k' is 5 by default.
- **Link-based representation**: "bag of documents" model with the document frequency equal to the number of incoming links associated with a, pointing to that document.

For document d,

- **Content-based representation**: "bag of words" model on page content. Each element records term frequency.
- **Link-based representation**: "bag of document" model with the document frequency equal to the number of incoming links pointing from that document.

In this way, we represent each a or d by a term/document frequency vector θ. We then normalize θ, dividing each element by the sum of all elements. Since θ may be sparse, we use Laplacian smoothing (esp. add-one smoothing) in practice, i.e., $\theta'_w = (\theta_w + \alpha)/(\sum_{w' \in \mathcal{V}} \theta_{w'} + \alpha|\mathcal{V}|)$, where $\alpha = 1$ and $\mathcal{V}$ is the vocabulary. We use both representations, and combine them by the geometric mean over similarities based on each type of representation.

Table 3: Representation ablation. Ranking performance NDCG@20 on TREC 2010 as a test set.

Methods	Content+Link	Link	Content
AAMSC	0.133	0.131 (-1.6%)	0.132 (-0.8%)
LinkProb	0.135	0.133 (-1.5%)	0.135 (-0.0%)
Combined+Max	0.133	0.132 (-0.8%)	0.133 (-0.0%)
SiteProbEx	0.132	0.131 (-0.8%)	0.131 (-0.8%)
M-ORG-RALM	0.135	0.132 (-2.3%)	0.134 (-0.8%)

We next present how to compute similarity matrices $\mathbf{W}^{(aa)}$, $\mathbf{w}^{(ad)}$ and $\mathbf{W}^{(dd)}$. For anchor-doc similarity $\mathbf{w}^{(ad)}$, we compute cosine similarity. Given anchor text line a and target page d, $\mathbf{w}_d^{(ad)}(a)$ is defined as $\theta'_d \cdot \theta'_a$.

For $\mathbf{W}^{(aa)}$ and $\mathbf{W}^{(dd)}$, we use the heat kernel (Lafferty and Lebanon 2005) to measure the affinity between multinomial distributions θ_i and θ_j, as it has been successfully applied in many IR and classification tasks (Diaz and Metzler 2007; Dillon and Collins-Thompson 2010). It is given by:

$$
\mathcal{K}(\theta_i, \theta_j) = \exp(-\frac{1}{t}\arccos^2(\sum_w \sqrt{\theta_{i,w}} \cdot \sqrt{\theta_{j,w}})) \qquad (9)
$$

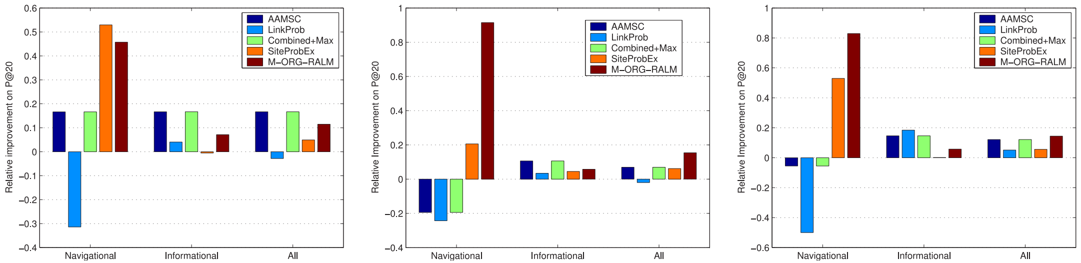
where t is a parameter controlling the decay of heat flow. We set it to 0.5. w is word/document within vector θ_i and θ_j. For $\mathbf{W}^{(aa)}$, we compute $\mathcal{K}_{(aa)}$ between all paired anchor text lines. For $\mathbf{W}^{(dd)}$, we only compute $\mathcal{K}_{(dd)}$ from each d to its k nearest neighbors with $k = 20$, i.e., from d to d' with $d' \in \mathcal{N}_d$. Using graph Laplacian, we choose to define $\mathbf{W}^{(aa)}$ ($\mathbf{W}^{(dd)}$) as the exponential of $\mathcal{K}_{(aa)}$'s ($\mathcal{K}_{(dd)}$'s) Laplace-Beltrami operator (Lafon 2004), given by:

$$
\begin{aligned}
\mathbf{W}^{(aa)} &= \exp(-l \cdot (I - \hat{D}_{(aa)}^{-1/2} \hat{\mathcal{K}}_{(aa)} \hat{D}_{(aa)}^{-1/2})) \qquad (10) \\
\mathbf{W}^{(dd)} &= \exp(-l \cdot (I - \hat{D}_{(dd)}^{-1/2} \hat{\mathcal{K}}_{(dd)} \hat{D}_{(dd)}^{-1/2})) \qquad (11)
\end{aligned}
$$

where $\hat{\mathcal{K}}_{(aa)}$ ($\hat{\mathcal{K}}_{(aa)}$) is the normalized affinity matrix with $\hat{\mathcal{K}}_{(aa)} = D_{(aa)}^{-1} \mathcal{K}_{(aa)} D_{(aa)}^{-1}$ and $\hat{\mathcal{K}}_{(dd)} = D_{(dd)}^{-1} \mathcal{K}_{(dd)} D_{(dd)}^{-1}$. $\hat{D}_{(aa)}$, $\hat{D}_{(dd)}$, $D_{(aa)}$, and $D_{(dd)}$ are diagonal matrices with $\hat{D}_{(aa)ii} = \sum_j \hat{\mathcal{K}}_{(aa)i,j}$, $\hat{D}_{(dd)ii} = \sum_j \hat{\mathcal{K}}_{(dd)i,j}$, $D_{(aa)ii} = \sum_j \mathcal{K}_{(aa)i,j}$, and $D_{(dd)ii} = \sum_j \mathcal{K}_{(dd)i,j}$ respectively. In this work, we always set parameter l to 0.1.

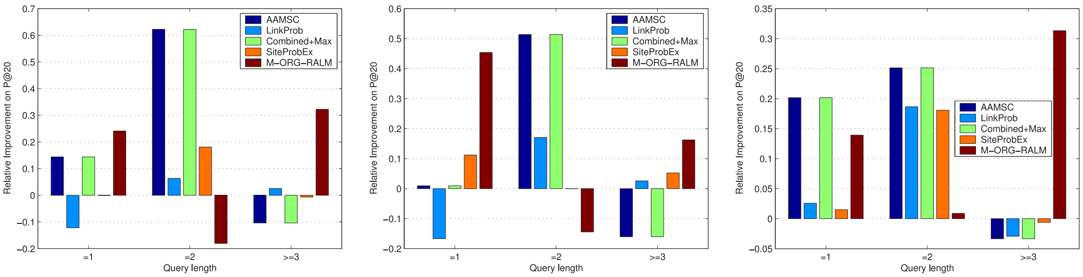
Experimental Setup

Data set and Evaluation. Our goal is to improve search relevance through incorporating anchor text context into estimating anchor text representation. The experiments are conducted on ClueWeb09 (Category B) data set. It includes 49.8M web pages and 940M hyperlinks, with 7.6M pages having in-coming links and 19M pages having enriched anchor text lines (Yi and Allan 2010; Metzler et al. 2009). Given that only a few pages associate with a large number of unique anchor text lines, we set the upper bound for the size of $\mathcal{A}_d$ to 200, keeping the most important anchor text lines (from basic estimators). The corpus is indexed through Indri Search Engine (Lemur Project 2010)



(a) S_{aa} (b) S_{ad} (c) $S_{aa'}$

Figure 2: Relative improvement (by using each constraint) over baseline anchor text importance estimators for P@20 on TREC 2009 as test set. We separate queries according to their intent. Topic 5, 15, 21, 23, 27, 31, 40, 41, 46 have navigational intent as revealed by manual inspection.



(a) S_{aa} (b) S_{ad} (c) $S_{aa'}$

Figure 3: Relative improvement (by using each constraint) over baseline anchor text importance estimators for P@20 on TREC 2009 as test set. We separate queries according to their length.

using Krovetz stemmer. We use the Ad hoc task of the TREC 2009 (50 queries) and 2010 (50 queries) Web track for evaluation. For each query, we first achieve its top 2000 search results by smoothed language model (Dirichlet smoothing) with parameter $\mu = 2000$, and then proceed reranking. Our main evaluation metrics are Precision (P@k) and Normalized Discounted Cumulative Gain (NDCG@k) (Jarvelin and Kekalainen 2000) at truncate level k. NDCG credits more when relevant results are ranked at higher positions.

Ranking Function. The way that anchor text representation influences document relevance estimation is through being part of document fields. Combining different document fields has been shown highly effective for retrieval on the web in previous work (Zaragoza et al. 2004). Representative ways include BM25F (Robertson, Zaragoza, and Taylor 2004) and fielded language models. In this work, we choose to use BM25F for showing the efficiency of combined fields without the loss of generality. BM25F is defined as:

$$BM25F(q,d) = \sum_{w \in q} \frac{\hat{t}f(w,d)}{k_1 + \hat{t}f(w,d)} \log \frac{N - df_w + 0.5}{N + 0.5} \qquad (12)$$

where N is the total number documents. df_q is the number of documents containing w. k_1 is fixed to 1000. $\hat{t}f(w,d)$ is the normalized term frequency weighted over all fields, given by

$$\hat{t}f(w,d) = \sum_{fi=\{anc,doc\}} wt(fi) \frac{tf(w,fi,d)}{1 + b_{fi}(\frac{l(fi,d)}{\bar{l}(fi)} - 1)} \qquad (13)$$

where $wt(fi)$ is trade-off between "anchor text" and "doc body". $tf(w,fi,d)$ is term frequency in field fi of document d. For anchor fields, $tf(w,anc,d)$ is the accumulated term frequency over all unique anchor text lines, weighted by the estimation of their relative importance $\hat{\mathbf{f}}_d$. $l(fi,d)$ is the length of field fi. $\bar{l}(fi)$ is average length of field fi over all documents. b_{fi} is fixed to 0.8. To estimate document relevancy, we learn parameters $wt(fi)$, ξ, μ, λ and ε (The latter four are from Equation 8) using coordinate ascent algorithm (Metzler and Croft 2007), driven to optimize mean average precision in training process.

Baseline Methods. The baseline anchor text importance estimation approaches are LinkProb (Dou et al. 2009), AAMSC (Fujii 2008), Combined+Max (Metzler et al. 2009), SiteProbEx (Dou et al. 2009), and M-ORG-RALM (Yi and Allan 2010). For each of these five approaches, we apply our regularization approach (denoted as CxtReg) and compare their ranking performance.

Experimental Results

We start our experiments by investigating the performance of our baseline models before and after we apply CxtReg. Performance is based on two-fold cross-validation, i.e., training on TREC 2009 (2010) queries and testing on TREC 2010 (2009) queries. We then conduct deeper analysis on the benefits from each constraint and representation.

Performance comparison. Table 1 and 2 show ranking performance on TREC 2009 and 2010 respectively, both as training and test set. CxtReg significantly enhances all base-

Table 1: Ranking performance on TREC 2009 of baseline anchor text importance estimators before and after we apply **CxtReg** on each one of them. Symbol §denotes statistically significant differences from baselines by a single-tailed student t-test.

	Baselines		As a Training Set		As a Test Set		Best Parameters
Methods	P@20	NDCG@20	P@20	NDCG@20	P@20	NDCG@20	$(\mu,\xi,\lambda,\varepsilon)$
AAMSC	0.350	0.296	0.390(+11.4%)§	0.303 (+4.5%)	0.351 (+0.2%)	0.290 (-0.3%)	(0,0.1,0,0.4)
LinkProb	0.359	0.290	0.393 (+9.8%)	0.307 (+5.8%)	0.382 (+6.4%)§	0.303 (+3.4%)	(0.1,0.8,0,0.2)
Combined+Max	0.350	0.280	0.392 (+12.0%)§	0.303 (+8.2%)§	0.351 (+0.2%)	0.296 (+5.7%)§	(0.2,0,0,0.4)
SiteProbEx	0.330	0.267	0.356 (+7.8%)	0.304 (+13.8%)§	0.351 (+6.3%)	0.293 (+9.7%)§	(0,0,0.8,0.1)
M-ORG-RALM	0.310	0.236	0.360 (+16.0%)§	0.304 (+28.8%)§	0.363 (+17.1%)§	0.301 (+27.0%)§	(0.5,0.8,0.1,0)

Table 2: Ranking performance on TREC 2010 of baseline anchor text importance estimators before and after we apply **CxtReg** on each one of them. Symbol §denotes statistically significant differences from baselines by a single-tailed student t-test.

	Baselines		As a Training Set		As a Test Set		Best Parameters
Methods	P@20	NDCG@20	P@20	NDCG@20	P@20	NDCG@20	$(\mu,\xi,\lambda,\varepsilon)$
AAMSC	0.241	0.129	0.268 (+10.8%)§	0.143 (+10.0%)§	0.264 (+9.4%)§	0.133 (+3.1%)	(0,0.2,0,1)
LinkProb	0.248	0.127	0.270 (+8.8%)§	0.141 (+11.0%)§	0.265 (+6.8%)	0.135 (+6.2%)§	(0.1,0.9,0.3,0.6)
Combined+Max	0.243	0.125	0.268 (+1.3%)	0.143 (+14.4%)§	0.264 (+8.8%)§	0.133 (+6.4%)§	(0,0.2,0,1)
SiteProbEx	0.265	0.133	0.269 (+1.3%)	0.135 (+1.5%)	0.272 (+2.4%)	0.132 (+0.8%)	(0.9,0.5,0.5,0.5)
M-ORG-RALM	0.257	0.129	0.262 (+1.9%)	0.134 (+3.8%)	0.257 (+0.0%)	0.135 (+5.1%)	(0.3,0.1,0.2,0.1)

line estimators on both data sets. This consistent superiority demonstrates the effectiveness of incorporating anchor text context on improving anchor text representation for retrieval. A closer look at their comparison suggests the following trends. First, the improvement is relatively independent with baseline estimators' details, but poorer baselines tend to benefit more from CxtReg. One possible reason is that the flexibility of leveraging the coverage and balance within anchor text representation help better respond information needs. Second, more diverse best parameter combinations tend to make CxtReg's relative improvement more diverse. For SiteProbEx and M-ORG-RALM, their best parameters learned from TREC 2009 and 2010 are more diverse, and their relative improvement on training and test sets is also more diverse. Assuming this two points are highly correlated, we infer that the benefits from CxtReg tends to be stable at least for some basic estimators, such as LinkProb and Combined+Max.

Benefits from individual constraint vs. query intent. Figure 2 shows the relative improvement when applying each individual constraint to queries with different intent. All three constraints consistently improve search relevance on informational queries, indicating incorporating anchor text context help more intelligently answer complex and rich information needs, rather than just locating home pages. It is especially useful for commercial search engines given that other ranking signals have been able to handle navigational queries well. Benefits on navigational queries depend on basic anchor text importance estimators.

Benefits from individual constraint vs. query length. Query length reflects how broad the information needs are. Figure 3 shows the the relative improvement when applying each individual constraint to queries with different length. Mostly three individual constraints outperform greater on queries with length being 2, indicating these constraints do not handle narrow information needs well. One possible reason is that the pages with narrow content attract few attention (i.e., anchors) from other pages, and CxtReg inevitably incorporates noise given that it draws from the whole anchor text context. The improvement on queries with one term is not stable, and highly depend on basic estimators. One may notice the abnormal properties shown on M-ORG-RALM. We conjecture that more anchor text lines incorporated by M-ORG-RALM improves the utilization of CxtReg.

Representation ablation. The efficacy of constraints relies on similarity estimation $\mathbf{W}^{(aa)}$, $\mathbf{w}^{(ad)}$, and $\mathbf{W}^{(dd)}$. We conduct representation ablation, shown in Table 3. Content-based representation outperforms link-based one. Their combination is superior to any one of them, indicating content-based and link-based representations contain complementary aspects.

Conclusion and Future Work

In this work, we propose to incorporate anchor text context into anchor text representation for improving search relevance. Three aspects include: (1) the relationship between anchor text lines with a same page, for estimating risk; (2) the relationship between anchor text representation and target pages; and (3) the relationship between similar anchors pointing to similar pages. We incorporate these three aspects into a unified optimization framework, aiming to enhance any basic anchor text importance estimators. Experimental results demonstrate our approach significantly improves all baseline estimators. Deeper analysis suggests our approach is especially useful for answering informational queries and bi-term queries. Representation enhancement helps further improve the efficacy of our approach.

The main limitation of this work is its efficiency issue, i.e., we only consider the top 200 important unique anchor text lines per page, and so how to mitigate this problem becomes part of our future work. In addition, a few interesting extensions include (1) designing document- and query-specific meta-features for controlling the relative importance of constraints; (2) applying the approach to other representation-based applications, such as cluster-based language models; and (3) designing more unified frameworks to optimize representation and retrieval models simultaneously. We hope to study these issues in the future.

Acknowledgments

The author is grateful to Professor Brian D. Davison for the guidance on her thesis work. This work was supported by a grant from the National Science Foundation under award IIS-0803605.

References

Amitay, E., and Paris, C. 2000. Automatically summarising Web sites — is there a way around it? In *Proceedings of the Ninth ACM International Conference on Information and Knowledge Management (CIKM)*.

Brin, S., and Page, L. 1998. The anatomy of a large-scale hypertextual Web search engine. In *Proc. of WWW*, 107–117.

Collins-Thompson, K. 2008. Estimating robust query models with convex optimization. In *NIPS*, 329–336.

Collins-Thompson, K. 2009. Reducing the risk of query expansion via robust constrained optimization. In *Proceedings of the 18th ACM conference on Information and knowledge management*, CIKM '09, 837–846. New York, NY, USA: ACM.

Craswell, N.; Hawking, D.; and Robertson, S. 2001. Effective site finding using link anchor information. In *Proceedings of ACM SIGIR*, 250–257.

Dai, N., and Davison, B. D. 2010. Mining anchor text trends for retrieval. In *Proc. of ECIR*, 127–139.

Dai, N.; Qi, X.; and Davison, B. D. 2011. Bridging link and query intent to enhance web search. In *Proceedings of the 22nd ACM conference on Hypertext and hypermedia*, HT '11, 17–26. New York, NY, USA: ACM.

Delbos, F., and Gilbert, J. C. 2003. Global linear convergence of an augmented Lagrangian algorithm for solving convex quadratic optimization problems. Research Report RR-5028, INRIA.

Diaz, F., and Metzler, D. 2007. Pseudo-aligned multilingual corpora. In *Proceedings of the 20th international joint conference on Artifical intelligence*, IJCAI'07, 2727–2732. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Dillon, J. V., and Collins-Thompson, K. 2010. A unified optimization framework for robust pseudo-relevance feedback algorithms. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10, 1069–1078. New York, NY, USA: ACM.

Dou, Z.; Song, R.; Nie, J.-Y.; and Wen, J.-R. 2009. Using anchor texts with their hyperlink structure for web search. In *Proceedings of SIGIR*, 227–234. ACM.

Eiron, N., and McCurley, K. S. 2003. Locality, hierarchy, and bidirectionality on the Web. In *Second Workshop on Algorithms and Models for the Web-Graph (WAW 2003)*. Extended Abstract.

Fujii, A. 2008. Modeling anchor text and classifying queries to enhance web document retrieval. In *Proceeding of the 17th international conference on World Wide Web*, 337–346. New York, NY, USA: ACM.

Gertz, E. M., and Wright, S. J. 2003. Object-oriented software for quadratic programming. *ACM Trans. Math. Softw.* 29:58–81.

Jarvelin, K., and Kekalainen, J. 2000. IR evaluation methods for retrieving highly relevant documents. In *Proc. of SIGIR*, 41–48.

Kraft, R., and Zien, J. 2004. Mining anchor text for query refinement. In *WWW '04: Proceedings of the 13th international conference on World Wide Web*, 666–674. New York, NY, USA: ACM.

Lafferty, J., and Lebanon, G. 2005. Diffusion kernels on statistical manifolds. *J. Mach. Learn. Res.* 6:129–163.

Lafon, S. 2004. *Diffusion Maps and Geometric Harmonics*. Ph.D. Dissertation, Yale University.

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

Lemur Project. 2010. Lemur project home page. `http://www.lemurproject.org/`.

Metzler, D., and Croft, B. 2007. Linear feature-based models for information retrieval. *Inf. Retr.* 10:257–274.

Metzler, D.; Novak, J.; Cui, H.; and Reddy, S. 2009. Building enriched document representations using aggregated anchor text. In *Proceedings of the 32nd SIGIR conference*, 219–226. ACM.

Murty, K. G. 1988. *Linear Complementarity, Linear and Nonlinear Programming*.

NIST. 2010. Text REtrieval Conference (TREC) home page. http://trec.nist.gov/.

Robertson, S.; Zaragoza, H.; and Taylor, M. 2004. Simple BM25 extension to multiple weighted fields. In *Proc. of CIKM*, 42–49.

Sarwar, B.; Karypis, G.; Konstan, J.; and Reidl, J. 2001. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*, WWW '01, 285–295. New York, NY, USA: ACM.

Wang, X., and Kankanhalli, M. 2010. Portfolio theory of multimedia fusion. In *Proceedings of the international conference on Multimedia*, MM '10, 723–726. New York, NY, USA: ACM.

Wang, J., and Zhu, J. 2009. Portfolio theory of information retrieval. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, 115–122. New York, NY, USA: ACM.

Wang, J. 2009. Mean-variance analysis: A new document ranking theory in information retrieval. In *Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval*, ECIR '09, 4–16. Berlin, Heidelberg: Springer-Verlag.

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

Yi, X., and Allan, J. 2010. A content based approach for discovering missing anchor text for web search. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, 427–434. New York, NY, USA: ACM.

Zaragoza, H.; Craswell, N.; Taylor, M. J.; Saria, S.; and Robertson, S. E. 2004. Microsoft cambridge at trec 13: Web and hard tracks. In *TREC*.

Zhu, J.; Wang, J.; Cox, I. J.; and Taylor, M. J. 2009. Risky business: modeling and exploiting uncertainty in information retrieval. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, 99–106. New York, NY, USA: ACM.