Incorporating Social Anchors for Ad Hoc Retrieval

Chia-Jung Lee
Center for Intelligent Information Retrieval
School of Computer Science
University of Massachusetts, Amherst
cjlee@cs.umass.edu

W. Bruce Croft
Center for Intelligent Information Retrieval
School of Computer Science
University of Massachusetts, Amherst
croft@cs.umass.edu

ABSTRACT

Anchor text has been widely used in web search as an effective complement to web page content. This motivates the investigation of similar sources of evidence about relevance. Social media postings often contain links to associated web pages, although typically not with anchor text. In this paper, we explore the use of these links and the text in the social postings as a form of anchor text (social anchors) for improving ad hoc search. Using a test collection based on ClueWeb09 together with associated social media, we show that by incorporating social anchor features, search effectiveness for "ad hoc" tasks can be significantly improved compared to state-of-the-art approaches. We also investigate the relative importance of social anchor features for retrieval, and show that query-dependent features are usually the key to better search performance.

Categories and Subject Descriptors

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

General Terms

Algorithms, Experimentation, Performance

Keywords

Social anchor, social text features, ad hoc retrieval

1. INTRODUCTION

The rich, dynamic, user-generated link information on the web is the basis of features such as anchor text and PageRank [4] that are a critical part of web search engine effectiveness. The use of anchor text has consistently been shown to improve effectiveness for web retrieval [8] [12] [26]. In the TREC "ad hoc" retrieval task, however, the effectiveness of anchor text and link analysis was less clear [15] [16]. Koolen and Kamps [18] suggested that the link sparsity problem and an insufficient collection size mainly accounted for the lack of effectiveness. To address the sparsity problem, Xing and Allan [30] proposed using content similarity between web pages, and discovered a web page's plausible missing in-link anchor text by using its most similar web pages' in-link anchor text. Others [14] argued that a good web collection of anchors need to have sufficient inter- and intra-server link densities. Recent results [1] [18] with the ClueWeb09 collection have shown that the use of anchor text can significantly improve retrieval effectiveness, indicating that the previous poor results were indeed due to collection properties such as collection size and link density.

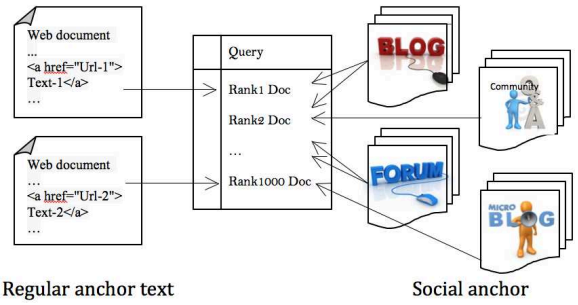

Figure 1: Collections of regular anchor text (left) and social anchors (right) for ranking documents in an ad hoc retrieval environment.

Given the effectiveness of anchor text as a feature in ranking functions, it is worth considering whether there are other sources of this type of information. The amount of social media generated has grown rapidly in recent years, and one interesting feature from our perspective is that many postings contain links to web pages, even though typical anchor text is not common. In this paper, we consider using these links and the content of the associated social postings as the basis of features for improving ad-hoc search. We define this type of social resource as **social anchors**. Specifically, we collect social anchors on the web that are restricted to the **social** domain. Social domains include those providing social discussions and communications that are usually presented via a variety of media such as blogs, forums, and microblogs. Accordingly, social anchors originate within these social domains and point to other web documents. Note that we are not looking at improving social media search, but instead using social anchors as a complementary source of evidence to standard anchor texts for ad-hoc search.

The right side of Figure 1 illustrates the idea of using so-

OAIR'13, May 22–24, 2013, Lisbon, Portugal.

Table 1: Statistics of returned results using 2000 random URLs as queries on different search APIs.

	Omgili	Twitter	Gblog	Gweb	Bing
#URLs	14	7	255	42	622
Avg. Ptr	5.71	23.57	14.82	4.9	65.94

cial anchors for ranking web documents in ClueWeb09. We collect a set of social anchors on the web that points to ClueWeb09 documents. With the social in-links, these documents can be enriched with external information including text descriptions and social signals from different domains such as blogs or forums. The left side of Figure 1, as a comparison, shows the typical use of a regular anchor text collection for relating the clickable text of web pages to the linked document.

In this paper, we conduct experiments to investigate how social anchors[1] can be useful for improving retrieval effectiveness. We also empirically compare the use of collections of regular anchor text and social anchors, and show that incorporating social anchors as new types of features into ranking algorithms can provide better results for finding evidence of relevance. This work is related to previous research that has expanded document representations using linked text [19], but focuses specifically on text from social media. We evaluate the proposed approach using TREC queries that are publicly available and "informational". We construct a set of social anchors associated with this query set based on which we demonstrate its significant potential for search effectiveness. While regular anchor texts had been shown to be effective for navigational queries in previous work, our results show that performance can be significantly improved using social anchors even for informational "ad hoc" queries.

The rest of this paper is organized as follows. In Section 2, we describe the approach taken to collect social anchors, based on which a set of potentially useful features can be extracted. In Section 3, we propose a feature-based linear model to incorporate the social anchor features for retrieval. In Section 4, we describe the evaluation results. We then discuss related work in Section 5 and conclude the paper in Section 6.

2. COLLECTING SOCIAL ANCHORS

In this section, we define and collect social anchors pointing to the ClueWeb09 documents, where each document d in ClueWeb09 corresponds to a web URL d_{url}.

While traditional media is more about "broadcasting", social media services are designed for conversations from interested users, encouraging voluntary contributions to useful information. Social communications are usually presented via a variety of media such as blogs, forums or microblogs. Accordingly, we refer to **social anchors** as those that originate within these **social domains** and point to other web documents.

Considering the entire web as the source corpus, an exhaustive search for social anchors to a document d would be infeasible for our experiments. Instead, we start by issuing each d_{url} as a query to a number of commercial search APIs. In this stage, we collect *all* possible returned results by using each API effectively as an inverted index. Different APIs allow different numbers of query requests and returned results. We iteratively send requests using different start and end offsets so as to collect the maximal results available. We should emphasize that we do not consider the ranks of the returned results and the possible experimental bias, if any, from the ranking algorithms of these commercial search engines is minimized. In the next step, we examine the general domain results in detail and remove documents that are not from social domains or do not exactly contain the query string d_{url}. After this filtering process, the set of remaining documents is regarded as **social anchor pages** pointing to d (or d_{url}). Since regular anchor texts are typically not available in these pages, we then construct **social anchor texts**, for the document d, using the title and surrounding texts of the social anchor pages.

[1]We focused mainly on blog data because of availability.

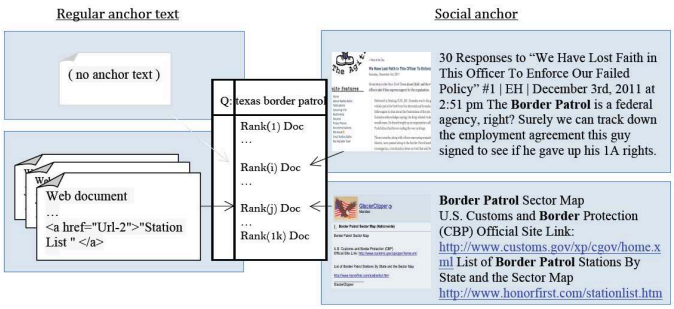

Figure 2: Examples showing the potential of incorporating social signals as new types of features into retrieval algorithm.

We start by testing several search APIs including the Omgili API, Twitter REST API, the Google Blog API, the Google Web Search API and the Bing Search API. To check whether sufficient incoming links could be obtained while keeping the search process efficient, we conduct a preliminary experiment to get an idea about the number of search results returned using 2000 seed URLs randomly selected from ClueWeb09. Table 1 shows the number of URLs out of the 2000 that have at least one result using the API, as well as the average number of returned results per d_{url}. Based on these statistics, we adopt the Bing Search API and the Google Blog API for collecting social anchors considering that the results from other APIs are too sparse.

We need to determine which domains are from social platforms. We therefore aggregate the domains out of the links returned from Google Blog API[2], the Wikipedia forum list[3] and the Wikipedia blog list[4]. This results in 20315 different domains and provides assurance that the filtered anchors are from social platforms.

Figure 2 shows, for the query topic "Texas border patrol", examples of social anchors (right) and typical anchor texts (left) pointing to ClueWeb09 documents within a ranked list. Anchor text has been shown to provide consistent effectiveness for retrieval, especially for navigational search. However, limitations such as the vague description "station list" that refers to border guard stations or missing anchor text makes it less useful for bringing relevant documents

[2]Google Blog API returns only blog documents.
[3]http://en.wikipedia.org/wiki/List_of_Internet_forums
[4]http://en.wikipedia.org/wiki/List_of_blogs

$\text{Doc}_{R(i)}$ and $\text{Doc}_{R(j)}$ to the top[5]. However, for the same $\text{Doc}_{R(i)}$ and $\text{Doc}_{R(j)}$, social anchors provide more precise information such as "The Border Patrol is a federal agency, right? Surely we can track down the employment agreement this guy signed ... " or "Border Patrol Sector Map" that matches the original query topic in terms of text descriptions. These examples[6] show that social platform users are motivated to provide more critical descriptions to certain topics. This presents the potential of incorporating social signals as new types of features into retrieval algorithm, where the effective auxiliary resource can be used to improve search performance.

2.1 Social Anchor Features

The next step is to extract useful evidence from the social anchors collected. The features based on social anchors can be classified into two types. The first type is query dependent that can be extracted from the social anchor texts. Previous research [12] suggests that counting the number of times that query terms match the anchor text may reflect the degree of relevance of the document to which the anchor page points. We propose four query dependent features from both micro and macro perspectives. A document d can have a number of social anchor texts $\{SA^i_{text}\}$. A macro view on the anchors combines all SA^i_{text} and renders a single virtual SA^c_{text}, whereas a micro view considers each SA^i_{text} individually.

- `MacroBin`: The feature computes the percentage of query terms existing in SA^c_{text}. The idea is to record whether or not each query term is included in SA^c_{text} (binary inclusion), and the feature is calculated by averaging the binary numbers over all terms within that query. This feature provides a simple yet effective estimation on the degree of vocabulary consistency. Common belief has it that the more matches between the anchor text and the query, the more relevant the document that the anchor points to.

- `MicroBin`: Similarly, the feature considers the binary existence of query terms in each SA^i_{text}, and aggregates the percentages by taking an average among all social anchors. The features of binary query terms inclusion can be effective even for documents having only a limited number of social anchors.

- `MacroLM`: The feature calculates the probability $P(Q|M_{SA^c_{text}})$ that the social anchor text's language model would generate the query terms. Language model features use the number of times query terms occur in social anchor texts, providing a numerical estimate compared to binary inclusion. Specifically, we adopt the definition of a unigram language model [25] for its simplicity and effectiveness.

- `MicroLM`: Based on the same idea, the language model feature $P(Q|M_{SA^i_{text}})$ is computed for each social anchor text, which are then combined into an average value.

[5]$\text{Doc}_{R(i)}$ and $\text{Doc}_{R(j)}$ are 2 relevant documents respectively ranked at positions i and j

[6]Note that the ranked list is a demonstrative example showing the same relevant documents receive different text information from different sources. The ranks in actual ranked lists will be different depending on the source exploited.

The second type of feature is independent of the content of the query, and are extracted from the properties of social anchor pages. Previous features such as PageRank [4] and HITS [17] estimate document priors based on link analysis without consideration of text content, and have demonstrated effective performance for large web collections. Similarly, we extract the link structure and time information from the social anchor pages.

- `NumAnchors`: Number of social anchors pointing to a document d. This in-degree feature can be interpreted as a measure of the popularity of d.

- `NumDomains`: Number of different social domains the social anchors come from for a document d. As suggested by Dou et al [10], anchor text from web sites of different domains should be considered as stronger evidence than that from the same site. We consider the diversity of social anchors by counting the number of different social domains from which the social anchors come. We hypothesize that the more different social domains associated with a document d, the more likely d is popular and more relevant to a topic.

- `ElapsedTime`: The average elapsed time for the social anchors pointing to a document d. Specifically, the elapsed time is computed by the time difference in the unit of years[7] between the creation time of the social anchor page and the current time[8]. Our hypothesis is that social discussions are usually time sensitive and certain query topics could be more popular than others within a specific period of time.

3. SOCIAL ANCHORS FOR RETRIEVAL

In this section, we first describe the incorporation of the social anchor features into a ranking formula based on a linear model. Then we detail the process of parameter learning for maximizing retrieval performance.

3.1 Feature-based Linear Model

Approaches that combine different evidence, text representations or search strategies for information retrieval have been shown to produce better effectiveness than a single system in most cases [5]. One of such approaches, the Markov Random Field for Information Retrieval (MRF-IR) [22], has consistently demonstrated state-of-the-art search performance by combining dependency text features. We thereby incorporate the social anchor features with the Sequential Dependency Model (SDM), an effective variant of MRF-IR.

Given a query q, the score of a document d in SDM [22] is computed as follows, where f_T, f_O and f_U are text feature functions associated with weight functions λ_T, λ_O and λ_U.

$$\begin{aligned} Score_{SDM}(q,d) = \lambda_T \textstyle\sum_{q_i \in q} log(f_T(q_i,d)) \\ +\lambda_O \textstyle\sum_{q_i \in q} log(f_O(q_i,q_{i+1},d)) \\ +\lambda_U \textstyle\sum_{q_i \in q} log(f_U(q_i,q_{i+1},d)) \end{aligned}$$

We briefly describe the feature functions: f_T measures the weight of unigram q_i in a document d, whereas f_O and f_U respectively compute the weights of exact phrase "$q_i q_{i+1}$" and $q_i q_{i+1}$ in an unordered window in a document d. Detailed descriptions can be found in [22].

[7]The use of "year" as the unit is because the social data gathered is much later than the ClueWeb09 crawl.

[8]This was 10/27/2011 for our experiments.

Based on the SDM framework, we can further combine the social anchor features using the following formula, where θ_j is the weight for the jth social anchor feature defined by the feature function ϕ_j.

$$Score(q,d) = \theta_{SDM} Score_{SDM} + \sum_j \theta_j \phi_j(q,d) \qquad (1)$$

By using linear models for combination, we have the advantage of computational simplicity, an interpretable model form, and the ability to compute certain diagnostic information about the quality of the fit. Consistent effectiveness in previous work [3] [22] also shows that linear models can be both efficient and effective.

3.2 Parameter Learning

With the linear model in Equation 1, our goal is to maximize a performance evaluation metric g by learning the set of free parameters $\Theta = \{\theta_{SDM}\} \cup \{\theta_j\}_{j=1}^{M}$ over the training set Q, where M denotes the number of social anchor features. Mathematically, the objective function optimizing g can be stated as follows,

$$\hat{\Theta} = \arg\max_{\Theta} g(Q, \{\pi(S_\Theta(D, Qi))\}_{i=1}^{N})$$

S_Θ, parameterized on Θ, denotes the ranking scores calculated based on Formula 1 for the query Q_i over the document corpora D. The permutation function π permutes collection documents in non-increasing order according to the ranking scores S_Θ. g can be optimized accordingly considering the ranked lists and the query set Q as input, varying on different selections of Θ.

Several learning approaches can be used for the purpose of parameter estimation. The use of the coordinate ascent algorithm proposed by Metzler and Croft [21] often converges to effective values [9] [3]. In this work, we adopt this algorithm for Θ estimation for both its simplicity and effectiveness; alternatives such as RankNet, RankBoost and AdaRank can also be used with competitive performance [9].

Largely used for unconstrained optimization problems, coordinate ascent iteratively maximizes a multivariate objective function by solving a sequence of scalar optimization subproblems. The algorithm searches over each coordinate axis and an ascent is made along one direction at a time. Specifically, it iteratively improves the estimate of the solution by maximizing over a selected coordinate $\theta' \in \Theta$ with all other $\Theta - \theta'$ fixed. The process is repeated until certain requirements are met, such as bounded gain compared to previous iteration. In this work, Mean Average Precision (MAP) is selected as the evaluation metric g. The parameters λ_T, λ_O and λ_U are set to 0.85, 0.15 and 0.05 according to consistent results in [22] [3]. The rest of the parameters Θ are estimated using coordinate ascent.

4. RETRIEVAL EXPERIMENTS

4.1 Experimental Setup

4.1.1 Dataset

We evaluate our approach using a large snapshot of the web – the ClueWeb09 Category B that contains a sample of 50 million English web documents of the entire ClueWeb09 crawl. The total 100 queries from TREC 2009 and TREC

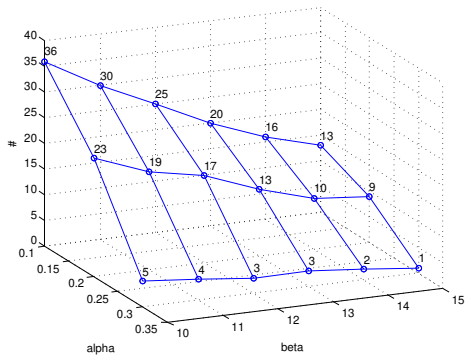

Figure 3: Number of topics based on different (α, β)

Table 2: Statistics of social anchor and all anchor for the initially retrieved document sets.

	Cnt	Inlink	Rel(%)	NRel (%)	NoJudge(%)
social	512.52	13.48	6.27	11.46	82.27
general	609.97	39.40	6.00	11.16	82.84

2010 Web track are considered in the experiments. To obtain reasonable and representative evaluation results, we need to assure a sufficient number of social anchors linking to ClueWeb09 documents. For a given query topic, we need to as well maintain a balance between the number of relevant and non-relevant documents having social anchors collected, thereby avoiding experimental bias if only relevant documents have social anchors. We construct two experimental parameters to address these concerns:

1. In a list of retrieved documents, the percentage of the sets of relevant and non-relevant documents having social anchors, denoted as Doc^s_{rel} and Doc^s_{nrel}, should exceed a threshold α. In this paper, we use results from the sequential dependency model `SDM` [22] to provide the list of retrieved documents.
2. The average number of social anchors among Doc^s_{rel} and Doc^s_{nrel} should be higher than a predefined value β respectively.

Figure 3 shows the number of qualified topics according to different choices of (α, β). In order to keep the test dataset associated with sufficient social anchors and balanced between Doc^s_{rel} and Doc^s_{nrel}, we select to use $(\alpha, \beta) = (0.1, 10)$. This indicate that at least 10% of relevant and non-relevant retrieved documents have social anchors, and the average numbers of both are at least 10. Accordingly, the final set of 36 query topics from the 100 is adopted for evaluation in our experiments. Given the difficulty of collecting social anchor data for TREC data set, the size of the resulting query set is relatively small. However, it is large enough to demonstrate the potential of our approach at least for informational queries and has the advantage of being publicly available. We plan to create a larger testbed in future work.

4.1.2 Retrieval setup

The document corpus is indexed using the open-source toolkit Indri[9], and the documents are stemmed with the

[9]http://lemurproject.org/indri/

Porter stemmer. The top 1000 documents are retrieved for evaluation, using a variety of measurement metrics including mean average precision (MAP), precision at certain positions (p@k), normalized discounted cumulative gain at certain positions (n@k) and mean reciprocal rank (MRR).

We compare our feature-based approach with several competitive methods. First, the sequential dependency model SDM [22] has consistently demonstrated state-of-the-art effectiveness, and is regarded as a strong baseline in this paper. Moreover, we extract standard anchor texts as in previous work from ClueWeb09 Category B using the harvestlinks method supported by Indri. Based on the standard anchor texts, we build an additional index $\texttt{IDX}_\texttt{A}$ that combines the anchor texts into document representations, based on which we examine the impact of standard anchor text. Finally, web search techniques using link analysis such as PageRank or HITS are commonly used. Accordingly, we compute the PageRank score in ClueWeb09 Category B as document priors that can be linearly combined into SDM. For all evaluations, we perform the 2-tailed t-test compared to baseline method SDM, and * is marked if p-value < 0.05.

Our evaluation on the linear model using social anchors is conducted by first retrieving the top 1000 documents using the SDM model for each $\{Q_i\}_{i=1}^N$. Based on these initial document sets, we rerank the results using Equation 1, where the set of parameters are learned using coordinate ascent. To avoid overfitting, we perform 9-fold cross-validation with 4 topics in each fold and report the average value. We conduct a range of experiments using different subsets of the social anchor features, and SDM+F denotes the run with all features adopted. One issue regarding using SDM to construct the initial list is that the ClueWeb09 collection is known to contain many spam documents. We address the problem by applying Waterloo's spam classifier [7] to the collection where documents with percentile-score< 60 are identified as spams. We denote the runs with spam filtering applied as $\texttt{SDM}_s$ and $\texttt{SDM}_s$+F that serve as a reference comparison.

Recall from Section 2.1 that we have an anchor collection from all general domains. As a comparison, we conduct an additional run that applied the proposed methodology to the general anchor collection and we denote this run as SDM+F(gen). The use of general domain documents lets us compare results and identify the potential utility of social signals. We summarize the statistics of social and general anchors in the initial document sets[10] in Table 2. The anchor count (Cnt) shows the average number of documents out of the 1000 that have anchors collected and the average number of inlinks (Inlink). Based on Cnt, we further calculate, for each $\{Q_i\}_{i=1}^N$, the percentage of relevant, non-relevant and no-judged documents, and take an average over the entire set Q. Table 2 shows that there is no bias towards relevant documents. Note that we use SDM+F(gen) only for the purpose of comparison with SDM+F, and thus all other experiments are conducted based on social anchors in the following sections.

4.2 Performance using Social Anchor Text – Query Dependent Features

To examine the effectiveness of individual features, we first incorporate one query dependent feature f at a time into the SDM model. The evaluation results are shown in Table 3.

[10]These statistics hold for other runs since reranking does not change the document sets.

Table 3: Retrieval performance using SDM and SDM+f.

model	MAP	p@5	p@10	n@5	n@10	MRR
SDM	.1699	.2222	.2638	.1219	.1561	.2632
MacroBin	.1789*	.2888*	.3083*	.1775*	.2025*	.3887*
MicroBin	.1775	.2722*	.2916*	.1617*	.1983*	.3686
MacroLM	.1706	.2277	.2611	.1246	.1605	.2755
MicroLM	.1700	.2277	.2611	.1190	.1568	.2617

Table 4: Retrieval performance using SDM and SDM+f.

model	MAP	p@5	p@10	n@5	n@10	MRR
SDM	.1699	.2222	.2638	.1219	.1561	.2632
NumAnchor	.1671	.2278	.2750	.1322	.1659	.2824
NumDomain	.1700	.2500*	.2611	.1464*	.1727*	.3254*
ElapsedTime	.1690	.2167	.2611	.1239	.1591	.2633

We can see that the incorporation of the binary inclusion features MacroBin and MicroBin resulted in significant improvements compared to the standard SDM. In particular, the evaluation metrics that measure the top retrieved results are significantly boosted, which is important to retrieval environments such as web search. We believe the improvements mainly come from the precise vocabulary generated by web authors in some popular forums or blogs. For the features based on language modeling techniques, MacroLM and MicroLM do not give as much improvement. This may be because the length of the social anchor texts is determined by the title plus the surrounding text[11], and thus the appearance of query words in social anchor texts is de-emphasized. Despite the limited improvements, we still keep the features because certain metrics, such as n@10 or MRR, can be improved in some runs. Moreover, comparing the performance between Macro- and Micro- features, the Macro- features are slightly better than the Micro- features but no significance difference was found.

In addition to the parameters learned by coordinate ascent, we are interested in the impact of each feature ϕ_j when the weighting function θ_j increases. To this end, we exploit fixed weighting method (FW) by varying θ_j from 0.1 to 0.9 by step of 0.1 with $\theta_{SDM} = 1 - \theta_j$, and show the results in the first row of Figure 4. We observe a common tendency that, when θ_j is increasing, the metrics that emphasize the top retrieved documents tend to get higher until some peak points are met, and decrease due to over-stressing the particular feature. Meanwhile, the metric MAP exhibits an inverse trend where increasing the weighting for anchor features usually results in a performance drop. This observation is consistent with Koolen and Kamps [18] who reported that anchor text is very effective for early precision but not for MAP.

4.3 Performance using Social Anchor Page – Query Independent Features

Another type of feature is the query independent features generated based solely on social anchor pages. We report the impacts of the combining each individual feature f with the SDM feature in Table 4. The NumAnchor feature, which considers the authority of the web documents, unexpectedly did not produce large improvements in general, though the metrics based on early precision are en-

[11]The window size is 20.

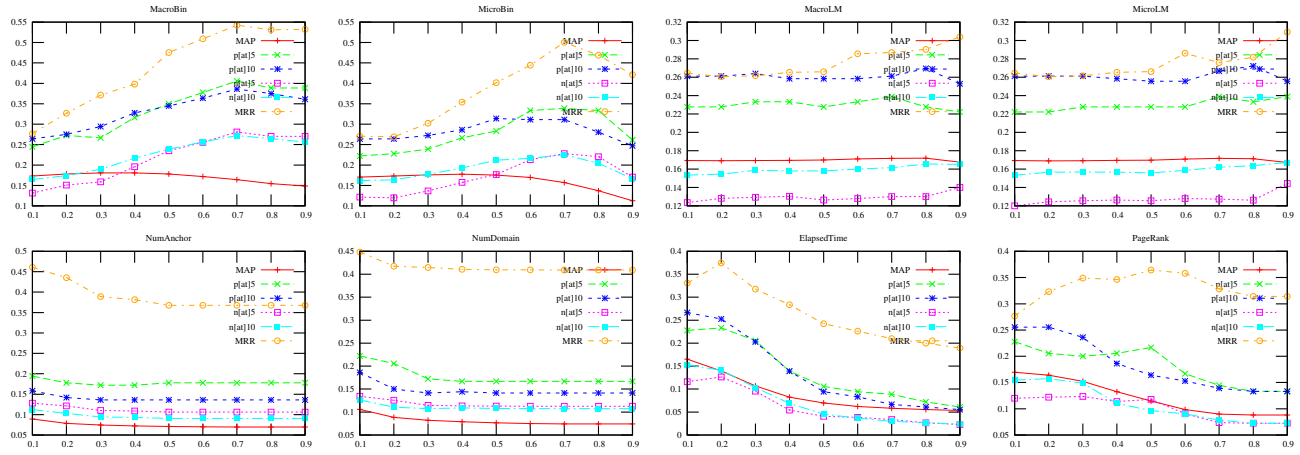

Figure 4: The performance trend by varying the weight for the social anchor feature f presented in x-axis.

hanced to some extent. This finding is again consistent with [18]. Koolen and Kamps mentioned that link evidence such as in-degree, whose main use is to separate the authoritative pages from less popular ones, does not contribute to effectiveness because of the characteristics of ClueWeb09 documents. Specifically, the ClueWeb crawl is based on a PageRank/POIC policy as opposed to typical breadth-first search [13]; therefore, ClueWeb09 documents generally have comparable high importance compared to previous collections. The incorporation of `NumDomain` can improve search performance especially for early precisions. This result implies that the diversity of social anchor pages should be considered for better modeling, backing up the findings by Dou et al [10]. The result of using `ElapsedTime` in the coordinate ascent runs is not as effective; we keep this feature for its potential in improving n@5 and n@10.

The second row of Figure 4 displays the impacts of different degrees of fixed weighting in the ranking formula for each query independent feature. A general observation can be drawn here that increasing the weighting functions for these features decreases all performance metrics in most cases. For comparison, we also plot the performance trend for the PageRank feature, which shows very similar results. The incorporation of content-free features, as a result, should be carefully dealt with since the performance drops quickly if they are over-emphasized.

In the following sections, unless specifically stated, we will use the coordinate ascent approach on the 9-fold cross-validation data rather than the fixed weight method.

4.4 Integrated Performance

We combine both query dependent and independent features with `SDM`, and show the integrated retrieval results in Table 5. We start by introducing the runs appearing in Table 5, where `QLM` represents the standard approach using query likelihood language modeling. $\texttt{SDM,IDX}_\texttt{A}$ shows the performance on the ClueWeb09 Cat B index combined with anchor texts. $\texttt{SDM+PR}_\texttt{FW}$ demonstrates the results of incorporating the PageRank feature using fixed weights[12], whereas the weighting functions are learned using coordinate ascent in $\texttt{SDM+PR}_\texttt{CA}$. To examine the impacts of different sets of social anchor features, `SDM+QDep` and `SDM+QIndep` respectively take into account query dependent and independent features for retrieval. `SDM+F` represents the retrieval effectiveness of using all 7 features while `SDM+F'` is performed after feature selection. `SDM+F(gen)` shows results using features extracted from general domain anchors. Finally, we can further integrate the PageRank feature together with the social anchor features, and denote the runs as `SDM+F+Pr` and `SDM+F'+Pr`, performed respectively before and after feature selection.

From Table 5, we can see the baseline used in the paper is very effective compared to `QLM`. For the $\texttt{SDM,IDX}_\texttt{A}$ run, while we do not see a significant boost on MAP as reported in [1] [18], different system implementations[13], ranking methods or query sets can account for this. However, we find

Table 5: Retrieval performance using `SDM` as baseline, a range of previous proposed methods, and the integrated runs using different features.

model	MAP	p@5	p@10	n@5	n@10	MRR
`QLM`	.1578	.1944	.2222	.1184	.1432	.2431
`SDM`	.1699	.2222	.2638	.1219	.1561	.2632
$\texttt{SDM,IDX}_\texttt{A}$	.1604	.2333	.2472	.1442*	.1543	.2854
$\texttt{SDM+PR}_\texttt{FW}$	.1696	.2278	.2556	.1199	.1548	.2767
$\texttt{SDM+PR}_\texttt{CA}$	.1700	.2278	.2611	.1201	.1539	.2622
`SDM+QDep`	.1779*	.2778*	.3083*	.1744*	.2051*	.3829*
`SDM+QIndep`	.1698	.2556*	.2667	.1468*	.1729	.3359*
`SDM+F(gen)`	.1721	.2667*	.2861	.1727*	.1894	.3807*
`SDM+F`	.1799*	.3000*	.3167*	.2064*	.2297*	.4238*
`SDM+F'`	.1816*	.3167*	.3278*	.2124*	.2316*	.4176*
`SDM+F+Pr`	.1797*	.3000*	.3278*	.2094*	.2372*	.4471*
`SDM+F'+Pr`	.1811*	.3333*	.3278*	.2199*	.2354*	.4232*

Table 6: Retrieval performance with spam filtering applied.

model	MAP	p@5	p@10	n@5	n@10	MRR
$\texttt{SDM}_s$	.1587	.3389	.3306	.2310	.2327	.4621
$\texttt{SDM}_s$`+F`	.1663	.4056*	.3667*	.2900*	.2681*	.5612*
$\texttt{SDM}_s$`+F'`	.1798*	.4389*	.4056*	.3132*	.2936*	.6106*

[12]The best performance is reported among $\theta_j \in \{0.1, ..., 0.9\}$.

[13]In [18], separate indexes were built for texts and standard anchor texts.

significant improvements on early precision measurements such as p@5, n@5 and MRR for this run. For the runs SDM+PR$_{FW}$ and SDM+PR$_{CA}$ using the PageRank feature, we find there are improvements in some cases but the differences are not significant.

Comparing the effectiveness of using the query dependent and independent features, SDM+QDep consistently outperforms SDM+QIndep. Generally speaking, we find that the social anchor features, either query dependent or independent, can be more effective than using single features such as PageRank or the combined index IDX$_A$.

To test the utility of social signals, we compare the runs SDM+F(gen) and SDM+F where the only difference is the corpus used for extracting features. The results suggest that using only social anchors for ranking can be more effective than the run that considers all general domains, validating the assumption that social resources could provide important evidence about relevance.

The integrated run SDM+F based on all the social anchor features produced significantly improved effectiveness. We see that the 2 types of features can be mutually complementary, resulting in better performance than using either one of them alone. Since the feature space is reasonably small (i.e., 7 social anchor features), we conduct an exhaustive search on $2^7 = 128$ combinations of F, by considering whether to include or exclude each feature and permuting all possibilities. The SDM+F' run further improves the result of SDM+F, where the selected feature set F' = {MacroBin, MicroBin, MicroLM, NumDomain, ElapsedTime}.

Moreover, we compare the performance of SDM+F and SDM+F' to the results of SDM,IDX$_A$ and SDM+PR$_{CA}$. From Table 5, we conclude that the social anchors make a significant difference to effectiveness compared to using regular anchor text and regular link analysis measures. Such effectiveness is consistent for both early performance and MAP. We expect to obtain even better results by combining the PageRank feature with features in sets F and F'. The runs SDM+F+Pr and SDM+F'+Pr in Table 5 show that incorporating PageRank can indeed improve early performance and should be included if one focuses on the task of web search.

Finally, Table 6 shows the retrieval performance with spam filtering applied. Comparing Table 6 to the corresponding entries in Table 5, we find that although MAP is somewhat degraded, early performance such as P@10 and MRR increase significantly after the spam filter is applied, which is consistent with the results in [7]. More importantly, the improvement from using social anchor features is validated with and without spam filtering, showing that incorporating social signals can be effective across different experimental settings.

5. RELATED WORK

Different types of information associated with web documents have been studied for improving retrieval performance.

Anchor text: Eiron and McCurley [12] showed that anchor text resembles real queries in terms of term distribution and length, thereby bridging the information gap between query and document representations. Other results [8] [26] indicated that the integration of anchor text information can be especially useful for tasks of entry site findings. There are, however, some problems regarding the use of anchor text. Dou et al [10] showed that certain spamming-like behavior on the web can decrease search performance, and argued that inter-server anchors should be more important than intra-ones. Studies [23] [30] showed that search effectiveness heavily depends on the density of anchors, and indicated approaches based on anchor aggregation can well address the problem of link sparsity. Recent research [18] [1] demonstrated that the use of anchor text can be more effective for the currently largest test collection ClueWeb09.

Social annotations: Social tagging on online portals can be useful for describing web documents provided by web content readers. Bao et al [2] observed that the social annotations benefit web search by providing good summaries and estimated popularity. Yanbe et al [28] showed that social bookmarking can be more effective for web search compared to the link structure alone. Recent research results [24] [27] have as well shown similar conclusions.

Query logs: Commercial search engine logs [31] provide the information of how potentially relevant web documents are to the queries issued by web users. The use of click-through data [11] [6] has been shown to improve search performance based on implicit user feedback.

Document Prior: Previous findings, such as PageRank [4] or HITS [17], suggest that the hyperlink structure on the web provides important information and can be effectively modeled as document priors. Bendersky et al [3] further indicated, in addition to link analysis, that modeling the content of documents such as page layout and texts can be critical to retrieval performance.

Document Expansion: Approaches enriching document representations such as [19] [20] utilized cluster-based language models to improve retrieval effectiveness. Topic-based approaches [29] considered the utility of different topic models for information retrieval. Our approach, which can be viewed as a kind of document expansion, incorporates social signals into retrieval algorithms as opposed to aggregating language usages using the target corpus. In this paper, we focus on studying the anchor text aspect of social media resources.

6. CONCLUSIONS

In this paper, we incorporated social anchor features into the ranking function to improve ad hoc retrieval performance. We collected social anchors linking to web pages retrieved using TREC queries and extracted a number of query dependent/independent features. Evaluation results demonstrated that the integration of social anchor features based on a linear model can be very effective compared to state-of-art retrieval algorithms, showing that social anchors are good external resources for effective ranking of informational queries. More importantly, the use of social anchors makes a significant difference to effectiveness compared to using regular anchor text and regular link analysis features, and the effectiveness remains consistent both with and without spam filtering.

We conducted studies on the relative importance of the social anchor features, and found that query dependent features can be more important for retrieval performance than the independent ones. For query dependent features, results of using binary inclusion of query terms showed better effectiveness compared with language modeling. The query independent feature NumDomain had shown to be more effective than others of the same type. These features are complementary to each other; combining all social anchor

features produced more effective results than using each of them alone.

For future work, it will be important for us to generate a larger test collection with social anchors. This will involve a large crawl of both web data and social media from the same time period. We plan to extend the social domains by including other resources such as community question answering services and microblogs in addition to the media types of blogs and forums.

7. ACKNOWLEDGMENTS

This work was supported in part by the Center for Intelligent Information Retrieval, in part under subcontract #19-000208 from SRI International, prime contractor to DARPA contract #HR0011-12-C-0016 , and in part by NSF IIS-1160894. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

8. REFERENCES

[1] V. N. Anh and A. Moffat. The role of anchor text in clueweb09 retrieval. In *Proc. of TREC*, TREC'10, 2010.

[2] S. Bao, G. Xue, X. Wu, Y. Yu, B. Fei, and Z. Su. Optimizing web search using social annotations. In *Proc. of WWW*, WWW '07, pages 501–510, 2007.

[3] M. Bendersky, W. B. Croft, and Y. Diao. Quality-biased ranking of web documents. In *Proc. of WSDM*, WSDM '11, pages 95–104, 2011.

[4] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30:107–117, April 1998.

[5] W. Bruce Croft. Combining approaches to information retrieval. *Advances in Information Retrieval*, pages 1–36, 2000.

[6] O. Chapelle and Y. Zhang. A dynamic bayesian network click model for web search ranking. In *Proc. of WWW*, WWW '09, pages 1–10, 2009.

[7] G. V. Cormack, M. D. Smucker, and C. L. A. Clarke. Efficient and effective spam filtering and re-ranking for large web datasets. *CoRR*, abs/1004.5168, 2010.

[8] N. Craswell, D. Hawking, and S. Robertson. Effective site finding using link anchor information. In *Proc. of SIGIR*, SIGIR '01, pages 250–257, 2001.

[9] V. Dang and W. Bruce Croft. Feature selection for document ranking using best first search and coordinate ascent. *Inf. Retr.*, 10:257–274, June 2007.

[10] Z. Dou, R. Song, J.-Y. Nie, and J.-R. Wen. Using anchor texts with their hyperlink structure for web search. In *Proc. of SIGIR*, SIGIR '09, pages 227–234, 2009.

[11] Z. Dou, R. Song, X. Yuan, and J.-R. Wen. Are click-through data adequate for learning web search rankings? In *Proc. of CIKM*, CIKM '08, pages 73–82, 2008.

[12] N. Eiron and K. S. McCurley. Analysis of anchor text for web search. In *Proc. of SIGIR*, SIGIR '03, pages 459–460, 2003.

[13] D. Fetterly, N. Craswell, and V. Vinay. The impact of crawl policy on web search effectiveness. In *Proc. of SIGIR*, SIGIR '09, pages 580–587, 2009.

[14] C. Gurrin and A. F. Smeaton. Replicating web structure in small-scale test collections. *Inf. Retr.*, 7:239–263, September 2004.

[15] D. Hawking. Overview of the trec-9 web track. In *Proc. of TREC*, 2000.

[16] D. Hawking, E. Voorhees, N. Craswell, and P. Bailey. Overview of the trec-8 web track. in. In *Proc. of TREC*, pages 131–148, 2000.

[17] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46:604–632, September 1999.

[18] M. Koolen and J. Kamps. The importance of anchor text for ad hoc search revisited. In *Proc. of SIGIR*, SIGIR '10, pages 122–129, 2010.

[19] O. Kurland and L. Lee. Corpus structure, language models, and ad hoc information retrieval. In *SIGIR*, pages 194–201, 2004.

[20] X. Liu and W. B. Croft. Cluster-based retrieval using language models. In *Proc. of SIGIR*, SIGIR '04, pages 186–193, 2004.

[21] D. Metzler and W. Bruce Croft. Linear feature-based models for information retrieval. *Inf. Retr.*, 10:257–274, June 2007.

[22] D. Metzler and W. B. Croft. A markov random field model for term dependencies. In *Proc. of SIGIR*, SIGIR '05, pages 472–479, 2005.

[23] D. Metzler, J. Novak, H. Cui, and S. Reddy. Building enriched document representations using aggregated anchor text. In *Proc. of SIGIR*, SIGIR '09, pages 219–226, 2009.

[24] M. G. Noll and C. Meinel. The metadata triumvirate: Social annotations, anchor texts and search queries. In *Proc. of WI and and IAT*, pages 640–647, 2008.

[25] F. Song and W. B. Croft. A general language model for information retrieval. In *Proc. of CIKM*, CIKM '99, pages 316–321, 1999.

[26] W. K. T. Westerveld and D. Hiemstra. Retrieving web pages using content, links, urls and anchors. In *Tenth Text REtrieval Conference*, TREC '02, pages 663–672, 2002.

[27] S. Xu, S. Bao, Y. Cao, and Y. Yu. Using social annotations to improve language model for information retrieval. In *Proc. of CIKM*, CIKM '07, pages 1003–1006, 2007.

[28] Y. Yanbe, A. Jatowt, S. Nakamura, and K. Tanaka. Can social bookmarking enhance search in the web? In *Proc. of JCDL*, JCDL '07, pages 107–116, 2007.

[29] X. Yi and J. Allan. A comparative study of utilizing topic models for information retrieval. In *Proc. of ECIR*, ECIR '09, pages 29–41, 2009.

[30] X. Yi and J. Allan. A content based approach for discovering missing anchor text for web search. In *Proc. of SIGIR*, SIGIR '10, pages 427–434, 2010.

[31] Z. Zhuang and S. Cucerzan. Re-ranking search results using query logs. In *Proc. of CIKM*, CIKM '06, pages 860–861, 2006.