Using Anchor Texts with Their Hyperlink Structure for Web Search

Zhicheng Dou[1], Ruihua Song[1], Jian-Yun Nie[2], and Ji-Rong Wen[1]
[1]Microsoft Research Asia, No. 49 Zhichun Road, Beijing, China, 100190
[2]DIRO, University of Montreal, CP. 6128 succ. Centre-ville, Montreal, Qc. H3C 3J7 Canada
[1]{zhichdou,rsong,jrwen}@microsoft.com, [2]nie@iro.umontreal.ca

ABSTRACT

As a good complement to page content, anchor texts have been extensively used, and proven to be useful, in commercial search engines. However, anchor texts have been assumed to be independent, whether they come from the same Web site or not. Intuitively, an anchor text from unrelated Web sites should be considered as stronger evidence than that from the same site. This paper proposes two new methods to take into account the possible relationships between anchor texts. We consider two relationships in this paper: links from the same site and links from related sites. The importance assigned to the anchor texts in these two situations is discounted. Experimental results show that these two new models outperform the baseline model which assumes independence between hyperlinks.

Categories and Subject Descriptors

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

General Terms

Algorithms, Experimentation

Keywords

Anchor Text, Hyperlink Structure, Web Site Relationship

1. INTRODUCTION

Anchor texts in Web documents provide a short description of the target document. Although they are initially created to help users navigate from one page to another, they usually provide an additional and complementary description of the document contents. On the other hand, anchor texts also share similar characteristics with Web queries [6], for example, they are usually short and descriptive. They have a better chance to match user queries than the content words of a document. These two reasons have motivated the extensive utilization of anchor texts in commercial search engines [2]. Experiments have proven their capability of improving Web search effectiveness. For example, Craswell et al. [5] found that anchor texts are even more useful than page content for navigational queries.

Although many methods have been proposed to exploit anchor texts, we notice that these methods usually assume anchor texts to be independent. For example, the count of an anchor text from different Web pages is simply accumulated and considered directly as a relevance signal of the destination page with respect to a query matching the anchor text. While the count of an anchor text may reflect, to some extent, the degree of correspondence of the document to it, a simplistic accumulation may fail to account for different situations where a hyperlink is created. Below are some of such situations.

(1) Two links can come from a same Web site or from two Web pages that are copies on different mirror sites. The fact that the anchor text associated is duplicated does not mean that the anchor text is twice more important.

(2) Two links can come from two Web sites with cooperative relationships. These Web sites influence mutually and tend to have similar hyperlinks and anchor texts. The anchor texts from such sites should not be considered as independent evidence.

(3) Hyperlinks may be purposely created to boost the ranking of the destination page in Web search. This is often the case for spam links. Such anchor texts should not be assigned a high importance.

The above situations have been documented in many studies. However, no study has been carried out to take them into account when exploiting anchor texts. This paper tries to address the problem of relationships between the source pages of the anchor texts so as to adjust the weight of links. Two types of relationship will be considered: the source pages are from the same Web site; the source pages are from related sites. Two new models are proposed to consider these relationships: the site-independent model and the site relationship model. The site-independent model assumes that different hyperlinks coming from the same Web site are identical; while the site relationship model further considers the relationships between Web sites (including the relationship between source site and destination site, and the relationship between different sources sites). The weight assigned to an anchor text is adjusted accordingly. Our experimental results show that these two new models can outperform the baseline model, which uses hyperlinks as if they are independent. In addition, the site relationship model

SIGIR'09, July 19–23, 2009, Boston, Massachusetts, USA.

performs the best. This suggests that a finer consideration of relationships between source pages can better account for the true indication of relevance of anchor texts.

The rest of this paper is organized as follows. In Section 2, we discuss related work. In Section 3, we introduce some basic knowledge about anchor data and introduce the data collection used in this paper. In Section 4, we present a framework in which anchor data is used to provide relevance evidence to search. We then explore two anchor models in Section 5. We report experimental results in Section 6 and conclude our work in Section 7.

2. RELATED WORK

The use of anchor data has been widely explored in the areas of Web information retrieval and knowledge discovery.

Eiron and McCurley [6] presented a statistical study of the characteristics of anchor texts. They found that anchor texts resemble real-world user queries in terms of term distribution and length. They also revealed that anchor texts could provide more authoritative information to search than titles or content of the document. Craswell et al. [5] experimented the use of anchor texts for site finding. They treated all anchor texts to a destination page as a pseudo anchor document. Each anchor text in the document is weighted by its frequency in the pseudo document. They revealed that anchor texts are more useful than content words for navigational queries. Westerveld et al. [19] used a similar method, but with a language model [14] instead of the BM25 model [16]. Following these studies, we also use anchor texts as forming a pseudo document. However, we will assume some relationships between anchor texts instead of treating them as independent elements.

Fujii et al. [7] explored an anchor model, which further breaks anchor texts into terms . The weight of a term is estimated by considering the weight of each anchor text to the document as well as the weight of the term in the anchor text. In this paper, we focus on whole anchor texts and do not break them into terms.

Anchor texts have also been used in other Web search related areas. Fujii et al. [8] proposed a model to identify synonyms of query terms in anchor texts in order to expand queries. Kraft and Zien [11] mined anchor text to generate refinements or related terms for a query. They showed that anchor texts could produce refinement suggestions of higher quality than content words. Lee et al. [12, 3] proposed to use link distribution of anchor texts over destination pages to distinguish navigational queries from informational queries. Fujii [7] further improved this method by using the distribution of query terms to deal with the queries that do not match the whole anchor texts, but part of them. Lu et al. [13] considered anchor texts of links to the same Web page as parallel texts. This allowed them to extract a live translation dictionary for cross-language IR. Amitay and Paris [1] proposed to use the structure of hypertext to produce summaries of Web sites.

In none of the above studies, the structure of the hyperlinks have been taken into account when the associated anchor texts are used to improve Web search. The simplifying independence assumption will raise several problems in the situations that we mentioned in Section 1. For example, the importance of anchor texts of spam links will be boosted, so are those in the duplicated pages. Therefore, we cannot rely on the raw distribution of anchor texts as if they are truly sampled at random. The fact that some Web pages are strongly related does indicate that there could be some correlation between anchor texts from them, and that we should trust less the raw count of anchor texts in such a situation. This problem is the focus of this paper. We will propose methods to deal with possible relationships between source pages of hyperlinks.

Table 1: An example anchor in the Web page http://www.sigir2009.org/ calls/papers

Name	Value
HTML source	Submitted papers should be in the <a href = “http://www.acm.org/sigs/publications/proceedings-templates” > ACM Conference style </a>...
source page	http://www.sigir2009.org/calls/papers
source site	sigir2009.org
destination page	http://www.acm.org/sigs/publications/proceedings-templates
destination site	acm.org
anchor text	ACM Conference style

Table 2: Symbols used in this paper. To reduce space and void confusion, we add some prefix letters in the symbols: A: Anchor, P: Page, and S: Site.

Symbol	Description
APPages(a,d)	The pages which link to destination page d using anchor text a.
APSites(a,d)	The sites which have at least one page linking to d using a
PSrcSites(d)	The set of domains which have at least one page linking to d
ASrcPages(a)	The pages which use anchor text a to link to other pages
ASrcSites(a)	The domains which have at least one source page containing a
ADstPages(a)	The destination pages linked by a
SDstSites(s_s)	The sites linked by site s_s
SSrcSites(s_d)	The sites linking to site s_d
S2SDstPages(s_s, s_d)	The pages from site s_d that are linked by pages from site s_s

3. THE ANCHOR DATA

To avoid confusion, we first explain some basic definitions about anchor data in this section. Table 1 shows an example of raw anchor. Some definitions including source page, destination page, source domain, and destination domain, are described in the table. A source page usually links to one or more different destination pages using different anchor texts, and a destination page can also be linked by several source pages using different anchor texts. Note that in this paper <source page, destination page, anchor text> is assumed to be unique, i.e. multiple occurrences of an anchor text linking a source page to a destination page are counted only once. This strategy is used because of the fact that the absolute number of occurrences of an anchor text within a source page is not strong indication of its importance for the destination page. Some authors may include multiple hyperlinks with the same anchor text to the same destination page, while other authors would include only one. A stronger indication is the number of source pages (and their relationships) linking to the destination page, which is the factor which we focus on.

There are some publicly available Web collections built for research purposes. Unfortunately they are usually a small portion of the whole Web and links are often broken. To study the problem in a more realistic setting, we use a larger Web page collection used by a real search engine. It includes about 100 million Web pages from about 4 million Chinese Web sites. There are about 55 million unique anchor texts. All examples and experimental results are generated from this Web snapshot. This bigger Web page collection allows us to analyze more realistic problems behind the anchor data, and to test our methods in a more realistic setting.

For convenience, we define some symbols about anchor data that will be used in the remaining sections of this paper in Table 2.

4. FRAMEWORK

In this section, we will present an anchor based ranking framework. We follow the work done by Craswell et al. [5] and Westerveld et al. [19], and build an "anchor document" for a destination page. Given a destination page d, the anchor document contains all the unique anchor texts of d's incoming links, and each anchor text a_i is associated with a weight $f(a_i, d)$. The anchor text a_i is treated as a phrase and $f(a_i, d)$ is considered as its frequency (or importance) in the anchor document. The anchor document can be represented as follows:

$f(a_1, d)$ × anchor text 1
$f(a_2, d)$ × anchor text 2
...
$f(a_i, d)$ × anchor text i
...
$f(a_n, d)$ × anchor text n

$f(a, d)$ can be defined in different ways. The way to define it is of crucial importance. In the work of Craswell et al. [5] and Westerveld et al. [19], the $f(a, d)$ is simply set as the count of links to d with anchor text a (i.e. the number of pages linking it). For example, if 5,531 pages link to `http://china.nba.com/` with the anchor text "NBA", then the achor text "NBA" is assigned a frequency of 5,531 in the anchor document, even if all these links are from a same site. In this paper, we try to define better estimations of $f(a, d)$ to improve Web search ranking.

For the purpose of Web search, a good estimation of $f(a, d)$ should satisfy the following requirement: For a query q, the document d_1, which is more relevant than another document d_2, should also be ranked higher than d_2 according to the anchor document (Problem 1). Only when this condition is satisfied can we expect that the anchor document is useful for Web search. This condition is difficult to verify. A simpler requirement that we try to satisfy is as follows: if the query is exactly the anchor text a, the pages which are directly linked by the anchor text should be correctly ranked, i.e, more relevant results should be ranked higher than less relevant results (Problem 2). Other documents and other queries can be handled by a ranking model (for example the Okapi BM25 model or the language model).

Let $p(d|a)$ be the probability that document d is authoritative for anchor text a; $p(a)$ be the probability that anchor text a is used on the Web, and $p(a, d)$ be the probability that an anchor-document pair<a, d> is important on the Web. $f(a, d)$ should be approximately proportional to $p(a, d) = p(a) \cdot p(d|a)$ so that more authoritative documents could be ranked higher when the query is the same as anchor text a. Following this principle, we define the following general form of weighting function $f(a, d)$ for an anchor text a and a destination page d:

$$f(a, d) = c \cdot p(a, d) \propto p(a) \cdot p(d|a) \tag{1}$$

In Section 5, we will develop several methods to estimate $p(a, d)$. We calculate the anchor text weight $f(a, d)$ using $p(a, d)$ and a multiplier c (to generate integral values of $f(a, d)$), and then generate the anchor documents. We will discuss several ways to define c and $p(a, d)$ in the next section. Actually c is set as a constant in most current retrieval models using anchor texts because it is the same for any anchor and document pair and can be discarded in ranking. As an anchor document is constructed for a destination page, different ranking models, for example the probabilistic Okapi model [16] and the language model [14], can be used to index such anchor text and to perform retrieval based on them. In this paper, we used the same Okapi BM25 model as in [5] to process them.

5. MODELS

In this section, we present several models to estimate $p(a, d)$. We first introduce the link independent model which has been explored in previous work. Two new models, the site independent model and the site relationship model, are then developed in Section 5.2 and Section 5.3 .

In the following sections, we use symbol D to stand for the page corpus comprising all Web pages, use A to denote all anchor texts, and use S to denote all Web sites. Implicitly, $p(a, d) = p(a, D, d)$ because $p(a, d)$ is estimated solely based on D in this paper.

5.1 The link independent model (LinkProb)

The link independent model is a model used in the previous studies. It assumes that the links (or source pages) to a destination page are independent and they are of equal importance to support the destination page. Each existing link <d_s, d_t, a_i> contributes an equal weight $p(a_i, d_s, d_t)$ to the authority of destination page d_t. Therefore, the probability $p_l(a_i, d_s, d_t)$ is: $p_l(a_i, d_s, d_t) = \frac{1}{\sum_{a \in A, d \in D} |\text{APPages}(a,d)|}$. We then have:

$$p_l(a, D, d_t) = \sum_{d_s \in D} p_l(a, d_s, d_t) = \frac{|\text{APPages}(a, d_t)|}{\sum_{a \in A, d \in D} |\text{APPages}(a, d)|}$$

This value is in fact the percentage of *links* pointing to the destination page d_t with anchor text a on the Web. This model is used by Craswell et al. [5] and Westerveld et al.[19]. In this model, the probability $p_l(d_t|a)$ is defined as: $p_l(d_t|a) = p_l(d_t|a, D) = \frac{|\text{APPages}(a,d_t)|}{|\text{ASrcPages}(a)|}$. $p_l(d_t|a)$ is abbreviated as **LinkProb** in the remaining sections of this paper. In the example shown in Figure 1(a), there are four links with the anchor text a from three different sites. We have $p_l(d_1|a) = 0.75$ and $p_l(d_2|a) = 0.25$.

5.2 The site independent model (SiteProb)

The link independent model makes a strong assumption of independence between links. However, links from the same site usually have a strong relationship (correlation) because they could be generated by the same webmaster. Webmasters or document authors may create duplicated links to help

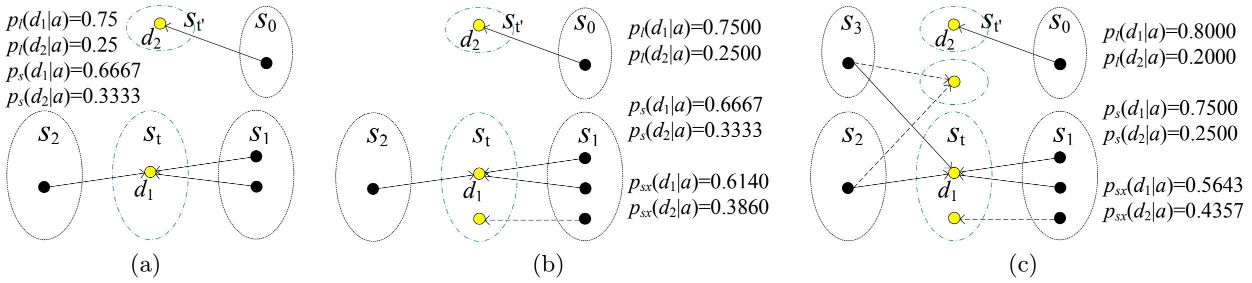


Figure 1: Examples for explaining the models. Links with anchor text a is drawn in solid arrow.

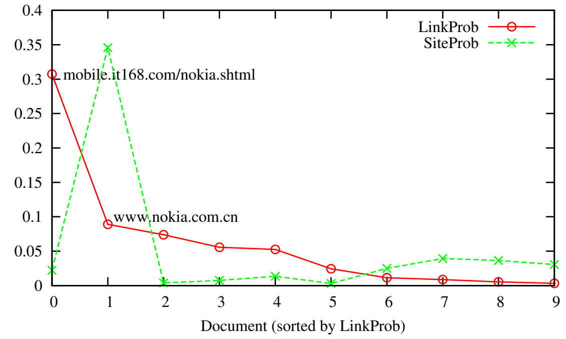


Figure 2: Pages linked by anchor text “Nokia”

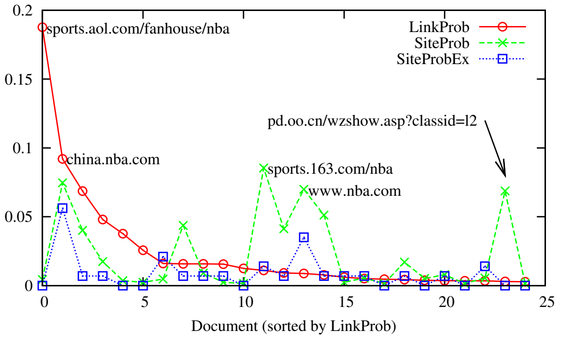


Figure 3: Pages linked by anchor text “NBA”

users navigate from different pages to the destination page, or to increase the link popularity of the destination page on purpose. In such cases, the link independent assumption does not hold and the probability of duplicated anchor texts is unduly increased.

To see the impact of using independent links, we plot the probability LinkProb ($p_l(d_t|a)$) using independent links for different pages for the anchor text “Nokia” in Figure 2 (see the red solid curve with circle point, LinkProb). We find the less authoritative page `http://mobile.it168.com/nokia.shtml` is ranked higher than the authoritative one `http://www.nokia.com.cn` (the homepage of Nokia Company). After analyzing the anchor data, we find that many sites contain multiple links with the anchor text “Nokia” linking to `http://mobile.it168.com/nokia.shtml`. For example, there are more than 100 links from the site `zhongsou.com` to this page, e.g. from the pags `http://bbs.zhongsou.com/sp/s/8331/1214567.html`, `http://bbs.zhongsou.com/sp/s/8331/1214704.html`, `http://bbs.zhongsou.com/sp/s/8331/1214717.html`, and `http://bbs.zhongsou.com/sp/s/8331/1214568.html`. These multiple links from the same site artificially increase the importance of the anchor text, while the real importance of it should be much lower.

To solve this problem, we make the following assumptions in our first approach – the site independent model: (1) The links with the same anchor text in the same site are strongly dependent and each site is allowed to give one vote on the authority of the destination page; (2) Different sites on the Web are independent and they are equally important to the destination page. Under these assumptions, we have the following definition of $p_s(a, s, dt)$:

$$p_s(a, s, d_t) = \text{constant} = \frac{1}{\sum_{a\in A,d\in D} |\text{APSites}(a, d)|} \qquad (2)$$

As $p_s(a, S, d_t) = \sum_{s\in S} p_s(a, s, d_t)$, and $p_s(a, s, d_t) = 0$ if none of pages in site s links to page d_t, we have: $p_s(a, S, d_t) = \sum_{s\in \text{APSites}(a,d_t)} p_s(a, s, d_t)$. Combining with Equation (2), we then have:

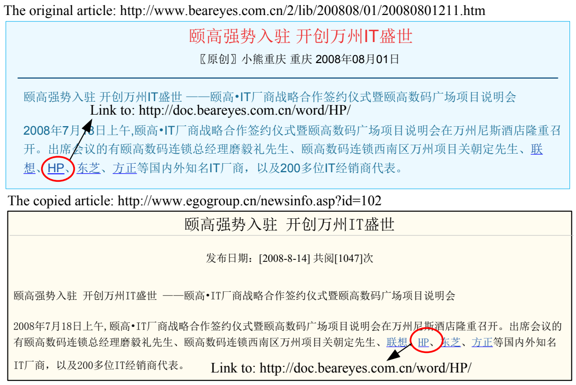
$$p_s(a, D, d_t) = p_s(a, D, d_t) = \frac{|\text{APSites}(a, d_t)|}{\sum_{a\in A,d\in D} |\text{APSites}(a, d)|}$$

and $p_s(d_t|a, D) = \frac{|\text{APSites}(a,d_t)|}{|\text{ASrcSites}(a)|}$. In this paper, $p_s(d_t|a, D)$ is abbreviated as **SiteProb**. In the example shown in Figure 1(a), $p_s(d_1|a) = 2/3$ and $p_s(d_2|a) = 1/3$ because there are 3 different source Web sites in total. Each of the links coming from site s_1 contributes a weight of 1/6.

The site independent model actually collapses all links with the same anchor text coming from the same domain. A large value of SiteProb is gained only if many sites link to this page using the anchor text. It is more difficult to spam than the link independent model because generating links in many different sites is much more difficult than generating links in one site. Figure 2 (see the green dashed curve with cross point, SiteProb) shows that for the query “Nokia”, the site independent model can successfully rank the authoritative page `http://www.nokia.com.cn` to the top.

5.3 The site relationship model (SiteProbEx)

The site independent model assumes that different Web sites are independent. However, Web sites can be strongly dependent. For example, there are mirror Web sites. The relationships between Web sites should also be taken into consideration. Figure 3 shows the top documents linked with the anchor text “NBA”. We see that the highly relevant document `http://china.nba.com` is not ranked at the top when the site independent model is used. By investigating the raw link data, we find that many links to less authoritative pages come from dependent Web sites. This strongly boosts incorrectly the unauthoritative pages, and penalizes the authoritative ones. The types of dependence between Web sites that we find frequently, which we consider in this paper, are introduced in Section 5.3.1 and Section 5.3.2.


The original article: http://www.beareyes.com.cn/2/lib/200808/01/20080801211.htm

颐高强势入驻 开创万州IT盛世

〖原创〗小熊重庆 重庆 2008年08月01日

颐高强势入驻 开创万州IT盛世 ——颐高•IT厂商战略合作签约仪式暨颐高数码广场项目说明会

Link to: http://doc.beareyes.com.cn/word/HP/

2008年7月18日上午,颐高•IT厂商战略合作签约仪式暨颐高数码广场项目说明会在万州尼斯酒店隆重召开。出席会议的有颐高数码连锁总经理磨毅礼先生、颐高数码连锁西南区万州项目关朝定先生、联想、HP、东芝、方正等国内外知名IT厂商，以及200多位IT经销商代表。

The copied article: http://www.egogroup.cn/newsinfo.asp?id=102

颐高强势入驻 开创万州IT盛世

发布日期：[2008-8-14] 共阅[1047]次

颐高强势入驻 开创万州IT盛世 ——颐高•IT厂商战略合作签约仪式暨颐高数码广场项目说明会

2008年7月18日上午, 颐高•IT厂商战略合作签约仪式暨颐高数码广场项目说明会在万州尼斯酒店隆重召开。出席会议的有颐高数码连锁总经理磨毅礼先生、颐高数码连锁西南区万州项目关朝定先生、联想、HP、东芝、方正等国内外知名IT厂商，以及200多位IT经销商代表。

Link to: http://doc.beareyes.com.cn/word/HP/

Figure 4: An example of copied pages

5.3.1 *The relationship between source site and destination site*

The link between two Web sites may not be as reliable as other links if the source site is dependent on the destination site. In this paper, we assume that the source site s_s is dependent on the destination site s_t if s_s links to many pages in s_t. Suppose S2SDstPages(s_s, s_t) defined in Table 2 is the set of pages that are in destination site s_t and linked by site s_s. We use $c(s_s, s_t)$ defined in Equation (3) to estimate the weight that s_s is dependent on s_t:

$$c(s_s, s_t) = \frac{1}{1 + \log |\text{S2SDstPages}(s_s, s_t)|} \tag{3}$$

A small value of $c(s_s, s_t)$ may be observed when s_s is a mirror site or cooperative site of s_t. In our data analysis, we found that a mirror site usually links back to its main domain with many links. A site may generate many links to pages in its cooperative site to boost their rankings. A large value of $c(s_s, s_t)$ can also be observed when s_t is a popular site and s_s is one of its users. The pages (e.g. news pages, articles, and blog posts) in a popular Web site are often copied and used by many small sites. If the original pages contain intra-site links and these links are preserved in the copies (See Figure 4 for example), the destination pages may receive many duplicated links from the copies on different sites.

In the example shown in Figure 1(b), two pages in site s_t are linked by site s_1 (note that the third link drawn in dashed line uses another anchor text). Using our calculation, we have: $c(s_1, s_t) = \frac{1}{1+\log 2} < 1$ while $c(s_2, s_t) = 1$ and $c(s_0, s_{t'}) = 1$.

5.3.2 *The relationship between source sites*

Links between Web sites may also be unreliable when source sites are dependent themselves because: (1) Links are duplicated when they are from mirror sites or copied pages; (2) Some sites are owned and designed by the same group of people to do search engine optimization (SEO) [1]. The links from these sites are less reliable; (3) Links can be added in some Web sites by other persons instead of the webmaster. For example, users can paste a list of links to the pages of forum or blog Web sites. In the Web page collection, we also find that many pages are attacked by hackers, and a large number of links are added in invisible blocks in these pages. For example, a large number of Web sites including `www.gilleda.com`, `www.gilleda.com`, `www.365job.com.cn`, `www.gdtravel.com.cn`, `www.shisu-lita.com`, and `www.90yi.com` are attacked, and their home pages are modified by inserting the following text, which generate a large number of hidden links [2]:

```
<marquee width=”8” height=”9” scrollamount=7881>
Site Links
<a href=http://www.sportblog.org.cn>sport blog</a>,
<a href=http://www.stockblog.org.cn>stock blog</a>,
<a href=http://www.gameblog.org.cn>game blog</a>,
<a href=http://www.wowgoldfood.cn>food</a>,
<a href=http://www.wowgoldflower.cn>flower</a>,
<a href=http://www.excamtest.cn>exam</a>, ...
```

In this paper, we simply assume that the source sites are dependent if the Web sites they link strongly overlap. For a specific page d, suppose PSrcSites(d) are the Web sites linking to d. SDstSites(s) is the set of Web sites pointed by site s. If sites PSrcSites(d) link to similar Web pages or Web sites, we reduce their weight for estimating the authority of destination pages. This is reasonable because: (1) The hyperlink structure including the out links of mirror sites are the same; (2) The same or a similar list of target Web pages are usually used by a same spammer. A possible reason is that the spammer usually wants to boost a large number of pages but he/she has limited resources (sources Web sites or human editors) to use. A simple way to spam is to insert all the links he/she wants to boost into all pages he/she has created, maintained, or attacked. Using our approach, these phenomena can be countered.

Since it is costly to calculate the relationship between two arbitrary Web sites, the probability that d_t is linked by a group of related sites is simplified to:

$$l(d_t) = \frac{\epsilon + \sum_{s \in \bigcup_{s_s \in \text{PSrcSites}(d_t)} \text{SDstSites}(s_s), s \neq s_t} idf(s)}{\epsilon + \sum_{s_s \in \text{PSrcSites}(d_t)} \sum_{s \in \text{SDstSites}(s_s), s \neq s_t} idf(s)}$$

Here $\bigcup_{s_s \in \text{PSrcSites}(d_t)} \text{SDstSites}(s_s)$ is the set of sites linked by the source sites of d_t. $\sum_{s_s \in \text{PSrcSites}(d_t)} \sum_{s \in \text{SDstSites}(s_s)} 1$ equals to the number of <site, site> pairs. s_t stands for the site of page d_t and it is excluded when $l(d_t)$ is calculated. We let $idf(s) = \log \frac{|S|+0.5}{|\text{SSrcSites}(s)|+0.5}$, a idf-like [17] formula, in order to reduce the negative impact of popular Web sites. We assume that a group of Web sites are strongly dependent only if the sites linked by them overlap and *are unpopular* (because popular Web sites may be normally linked by many Web sites together). ϵ is a smoothing parameter, and we let ϵ=10E-8 in this paper.

Suppose in Figure 1(b) and Figure 1(c), each idf value of the destination sites (drawn in dashed circle) equals to 1. In Figure 1(b), we will have $l(d_1) = \frac{\epsilon+0}{\epsilon+0} = 1$ and $l(d_2) = \frac{\epsilon+0}{\epsilon+0} = 1$. In Figure 1(c), we will have $l(d_1) = \frac{\epsilon+1}{\epsilon+0+1+1} \approx 0.5$ and $l(d_2) = \frac{\epsilon+0}{\epsilon+0} = 1$. We see that these values can reflect how unique a destination is linked and how important each link is to the destination page.

5.3.3 *The site relationship model*

The site relationship model considers the above relationships among Web sites. It assumes that different sites may have different weights for voting the authority of destination page, i.e., $p_{sx}(a, s, d_t) \neq constant$. Suppose $p^n(a, s, d_t)$ is the constant contribution of a normal link from site s to page d_t. We will add different weights to this normal contri-

[1] `http://en.wikipedia.org/wiki/Search_engine_optimization`

[2] We observed this in January 2009. You may fail to verify this as the content of these pages may have changed.

bution considering different relationships between Web sites by using the following equation:

$$p_{sx}(a, D, d_t) = \sum_{s \in \text{APSites}(a,d_t)} p^n(a, s, d_t) \cdot c(s, s_t) \cdot l(d_t)$$

Here s_t stands for the site of page d_t. $p_{sx}(a, D, d_t)$ and $p_{sx}(d_t|a, D)$ can be calculated once $p^n(a, s, d_t)$ is calculated.

$$p_{sx}(d_t|a, D) = \frac{l(d_t) \sum_{s_s \in \text{APSites}(a,d_t)} c(s_s, s_t)}{\sum_{d \in \text{ADstPages}(a)} l(d) \sum_{s_{s'} \in \text{APSites}(a,d)} c(s_{s'}, s_d)}$$

$p_{sx}(d_t|a, D)$ is abbreviated as **SiteProbEx**.

In Figure 1(b), $p_{sx}(d_1|a) = \frac{1\times(1+1/(1+\log 2))}{1\times1+1\times(1+1/(1+\log 2))} \approx 0.6140$ and $p_{sx}(d_2|a) \approx 0.3860$. In Figure 1(c), as $c(s_1, s_t) = \frac{1}{1+\log 2}$, $c(s_2, s_t) = 1$, $c(s_3, s_t) = 1$, $c(s_0, s_{t'}) = 1$, $l(d_1) = 0.5$, and $l(d_2) = 1$, we have:

$$p_{sx}(d_1|a) = \frac{0.5 \times (1 + 1 + 1/(1 + \log 2))}{0.5 \times (1 + 1 + 1/(1 + \log 2)) + 1 \times 1} \approx 0.5643$$

And $p_{sx}(d_2|a) = 0.4357$. Note that $p_{sx}(d_1|a)$ is not much bigger than $p_{sx}(d_2|a)$ despite the fact that it is linked by two more sites.

Figure 3 shows that the pages `http://china.nba.com` and `http://www.nab.com` can be successfully ranked to the top if the site relationship model is used. The page `http://sports.163.com/nba` is ranked to the top in the site independent model because the Web site `163.com` has many mirror sites or friend sites such as `http://www.netease.com` and `http://163.go24k.com` and they all contain links with anchor text "NBA" linking to `http://sports.163.com/nba`. Furthermore, `163.com` is a popular Web site in China and it produces many Web pages containing articles such as sport news. Some pages which contain intra-site links with anchor "NBA" are copied by other Web sites. As the site relationship model can account for the relationships between Web sites, the resulting vote for page `http://china.nba.com` and `http://www.nab.com` are corrected and these pages are successfully ranked to the top.

6. EXPERIMENTS

In Section 4, we stated that our goal is to generate more accurate weights for each anchor text and to improve the overall ranking base on the BM25 model. In this section, we will examine whether our models can generate better weights for documents directly linked by anchors, and then evaluate the ranking performance using the Okapi BM25 model.

6.1 Models validation on navigational queries

We use 1,131 navigational queries sampled from query logs to evaluate the accuracy of anchor data generated by different models. The relevance of linked documents for a given query is manually judged. Usually only one or a few perfect documents (for example mirror pages) are judged as relevant. For a given query, we sort its linked documents by their weights estimated by the three models, and then evaluate the ranking accuracy over all queries using the MRR (Mean Reciprocal Rank) metric [18]. We also use the Suc@k which means that percentage of queries for which at least one relevance result is ranked up to position k (including k).

Figure 5 shows the experimental results. We find that the site relationship model (SiteProbEx) significantly outperforms the site independent model (SIteProb) with a MRR gain of 0.07 (p<0.001) and a Suc@1 gain of 0.11, and the site independent model outperforms the link independent model with a MRR gain of 0.08 (p<0.001). Note that the Suc@10 differences between the three models are smaller than the Suc@1 differences.

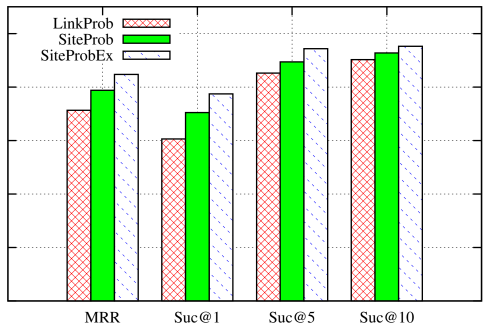


Figure 5: Results on 1,131 navigational queries

Table 3: Performance comparison of models by queries. The value in a cell means the number of queries on which the row model outperforms the column model (relevant document is ranked higher).

	LinkProb	SiteProb	SiteProbEx
LinkProb	/	48(4.24%)	59(5.21%)
SiteProb	267(23.60%)		71(6.28%)
SiteProbEx	323(28.56%)	183(16.18%)	/

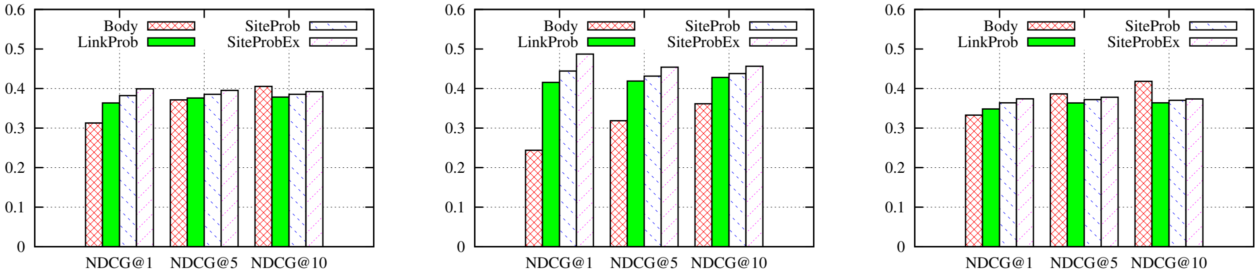
We also compare each two models by counting the number of improved queries. Table 3 shows the results. The value in a cell means the number and percentage of queries on which the row model outperforms the column model. We find that 23.6% of the queries are improved by the site independent model over the link independent model. The site relationship model further improves 16.18% of the queries over the site independent model. The site relationship could improve the accuracy of anchor text for 28.56% of navigational queries. This means that our models are useful to improve anchor texts for navigational queries.

6.2 Ranking experiments

In this section, we use the Okapi BM25 model to retrieve the anchor document generated by different models and evaluate the ranking accuracy. We use the same parameters (k_1 = 2.0, b = 0.75) as previous work [5].

We use a dataset which contains 3,000 randomly sampled queries and about 140 of returned documents for each query. The documents are manually judged by human editors. A five-grade (from 1 to 5 meaning from bad to perfect) rating is assigned for each document. There are both navigational and information queries in this dataset. We will experiment with the navigational queries (674 out of 3000) and informational queries together and separately.

The ranking accuracy is measured using a Normalized Discounted Cumulative Gain measure (NDCG) [9] based upon human judgments on test documents. We use the same configuration for NDCG as [4]. More specifically, for a given query q, the NDCG@K is computed as: $N_q = \frac{1}{M_q} \sum_{j=1}^{K} (2^{r(j)} - 1)/\log(1 + j)$. M_q is a normalization constant (ideal DCG) so that a perfect ordering would obtain NDCG of 1; and each $r(j)$ is a human rating of the result returned at position j. NDCG is well suited to Web



(a) All queries (b) Navigational queries (c) Informational queries

Figure 6: Ranking results using BM25 scoring over all, navigational, and informational queries

search evaluation, as it rewards relevant documents in the top ranked results more heavily than those ranked lower. We report NDCG@1, NDCG@5, and NDCG@10 in this paper.

6.2.1 *Performance of different models*

We first experiment with each model and compare their performances. We use the BM25 retrieval model to rank the documents based on the anchor documents generated by each model. To compare with the anchor models, we also experiment with the body text. Experimental results are shown in Figure 6. We can make the following observations:

(1) All three anchor models outperform the body model at NDCG@1 (p=1.53E-12, 4.87E-21, and 1.54E-29 correspondingly), while the body model outperforms the anchor models at NDCG@10 (p=7.26E-07, 3.38E-7, and 6.23E-6). This is reasonable because the quality of anchor text is usually high but its amount is limited, especially for tail documents. Figure 6(b) and Figure 6(c) show that anchor text is more useful for navigational queries than for informational queries. These observations are consistent with those in previous work [5, 6, 10].

(2) The site independent model and the site relationship model outperform the link independent model, especially for navigational queries at NDCG@1 (p <0.001 for all). This means that our new models can assign more reliable weights to anchor texts than the link independent model and help find good results at the top for navigational queries. Furthermore, the site relationship model performs the best. This means that we could further improve the quality of anchor by considering the relationships among Web sites.

6.2.2 *Combining body and anchor text*

Since the body text and anchor texts may reflect different aspects of a document, they are usually combined together in real-world use. We also test the approach combining body texts and anchor texts in this experiment. Robertson et al. [15] pointed out that a linear combination of BM25 scores is problematic and proposed a linear combination of term frequencies (BM25F). In this paper, we use the same term frequency combination method as in [15]. Suppose $w_{Body}(i,j)$ and $w_{Anchor}(i,j)$ are the term frequencies of term i in body and anchor text fields of document j. The BM25 score is calculated based on the aggregated term frequency $w(i,j)$ over body and anchor text fields as follows:

$$wi, j = \alpha \cdot w_{Body}(i,j) + (1-\alpha) \cdot w_{Anchor}(i,j)$$

where α is a combination parameter that decides the weight of body and anchor text in the aggregated BM25 score. The combined document length is also calculated using the same method. Figure 7 shows the experimental results when different settings of α are used. From this figure, we can observe the following facts:

(1) Figure 7(b) and Figure 7(e) show that the site relationship model (SiteProbEx) outperforms the site independent model and the link independent model for navigational queries when the same combination parameters are used (p <0.001 for all parameters). The site independent model outperforms the link independent model on the top 1 result of navigational queries when $\alpha \leq 0.7$. Figure 7(f) shows that the NDCG@10 differences for informational queries between each model are limited.

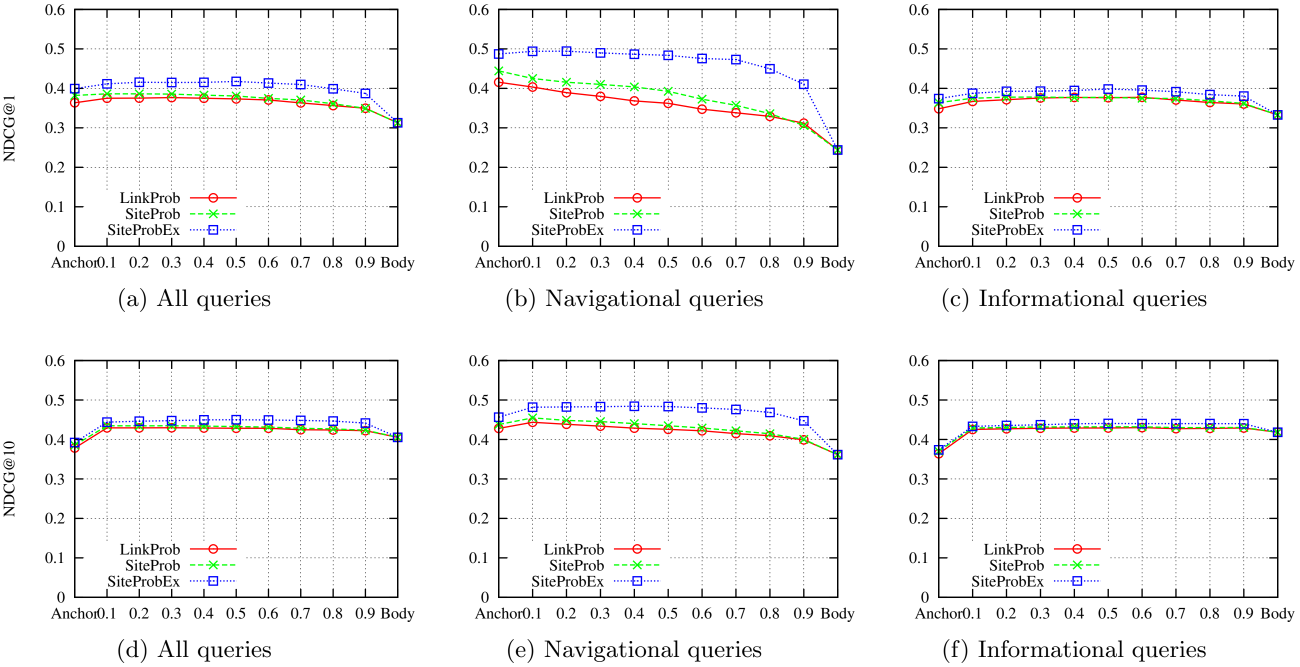
(2) Figure 7(b) shows that adding body text to anchor only slightly improves NDCG@1 for navigational queries. There are only about 0.012 NDCG@1 gain (p=6.69E-05) for the site relationship model ($\alpha = 0.1$). This suggests that anchor text is already indicative enough for navigational queries if only top 1 result is sought. However, body text is still useful to help improve the overall ranking of top 10 results.

(3) Figure 7(f) and Figure 7(c) show that the combination model outperforms both the body model and the anchor model for informational queries, especially on NDCG@10. This means that for informational queries, we can combine body text and anchor text together to yield a better ranking.

7. CONCLUSIONS

Although the use of anchor texts has been widely explored in both industrial and academic communities, anchor texts have been assumed to be independent. In real situations, this assumption does not hold. Links from the same Web site can be dependent, and links from related Web sites can also be dependent. In this paper, we examined several typical situations in which anchor texts can be dependent. Two new models are then proposed to arrive at a better estimation of the importance of anchor texts for a destination page, namely the site independent model and the site relationship model.

The site independent model supposes that different hyperlinks coming from the same Web site are strongly dependent. We consider them to be duplicates in this paper. Therefore, the link from one site is counted only once. This model avoids incorrect boosting of an anchor text when it is used many times on the same source site. The site relationship model further considers the relationships between Web sites. Links between related Web sites are considered to be less reliable and are assigned lower weights than links from unrelated Web sites. We analyzed some typical relationships between Web sites, and modeled their influence based on two



(a) All queries

(b) Navigational queries

(c) Informational queries

(d) All queries

(e) Navigational queries

(f) Informational queries

Figure 7: NDCG@1 (a, b, c) and NDCG@10 (d, e, f) results of combining body and anchor using BM25F with different settings of combination parameter α.

factors: the frequency that the source site links to the destination site, and the duplication degree of the sites linked by the source sites. Our experimental results show that using the two new methods proposed to consider anchor texts, we can further improve Web search rankings of the BM25 retrieval model, especially for the navigational queries.

In this paper, we focused on a limited number of dependence relations between Web sites, and many other types of relation still need to be investigated. The methods we proposed are based on strong assumptions. In our future work, we will further refine the models by making more realistic assumptions, on the one hand, and by incorporating more types of relation, on the other hand. Despite the limitation of our current work, our results already strongly indicate that the consideration of relationships between anchor texts is a promising direction to further improve Web search.

8. REFERENCES

[1] E. Amitay and C. Paris. Automatically summarising web sites: is there a way around it? In *Proceedings of CIKM '00*, pages 173–179. ACM, 2000.

[2] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of WWW '98*, pages 107–117, 1998.

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

[4] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender. Learning to rank using gradient descent. In *Proceedings of ICML '05*, pages 89–96, New York, NY, USA, 2005. ACM Press.

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

[6] N. Eiron and K. S. McCurley. Analysis of anchor text for web search. In *Proceedings of SIGIR '03*, pages 459–460, New York, NY, USA, 2003. ACM.

[7] A. Fujii. Modeling anchor text and classifying queries to enhance web document retrieval. In *Proceeding of WWW '08*, pages 337–346. ACM, 2008.

[8] A. Fujii, K. Itou, T. Akiba, and T. Ishikawa. Exploiting anchor text for the navigational web retrieval at ntcir-5. In *Proceedings of NTCIR-5 Workshop Meeting*, 2005.

[9] K. Järvelin and J. Kekäläinen. Ir evaluation methods for retrieving highly relevant documents. In *Proceedings of SIGIR '00*, pages 41–48, New York, NY, USA, 2000. ACM Press.

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

[11] R. Kraft and J. Zien. Mining anchor text for query refinement. In *Proceedings of WWW '04*, pages 666–674. ACM, 2004.

[12] U. Lee, Z. Liu, and J. Cho. Automatic identification of user goals in web search. In *Proceedings of WWW '05*, pages 391–400, New York, NY, USA, 2005. ACM Press.

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

[14] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *Proceedings of SIGIR '98*, pages 275–281. ACM, 1998.

[15] S. Robertson, H. Zaragoza, and M. Taylor. Simple bm25 extension to multiple weighted fields. In *Proceedings of CIKM '04*, pages 42–49. ACM, 2004.

[16] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-beaulieu, and M. Gatford. Okapi at trec-3. In *Proceedings of TREC-3*, pages 109–126, 1995.

[17] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage.*, 24(5):513–523, 1988.

[18] E. Voorhees. Trec-8 question answering track report. In *Proceedings of the 8th Text Retrieval Conference*, pages 77–82, 1999.

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