Navigation Retrieval with Site Anchor Text

Hideki Kawai Kenji Tateishi Toshikazu Fukushima
NEC Internet Systems Research Labs.
8916-47, Takayama-cho, Ikoma-city, Nara, JAPAN
{h-kawai@ab, k-tateishi@bq, t-fukushima@cj}.jp.nec.com

Abstract

In this paper we present an information retrieval system that indexes only site anchor text to verify the efficiency of reference information in a navigation retrieval task. We propose two relevancy measures to maximize limited information: reference consistency and specificity of word combination.

Our results show that navigation retrieval with a site anchor text can pinpoint highly relevant documents despite using one-thousandth less information than traditional full-text search systems.

Keywords: Site Anchor Text, Navigation Retrieval, Reference Consistency, Specificity of Word Combination

1 Introduction

A navigation retrieval task conducted in a NTCIR-4 WEB task B is defined as a "known item retrieval." It assumes that a user searches for one or more "representative Web pages" about an item with which the user is already familiar. However, the searcher does not necessarily know anything about the Web pages themselves. For this task, it is important to determine both relevancy and representativeness of the document.

Site anchor text is defined as the text in a link indicating the top page of a given Web site from external Web sites [1]. Site anchor text summarizes the content of the Web site, and the citation frequency indicates the representativeness of the Web site. We participated in such an NTCIR-WEB task to verify the efficiency of site anchor text in a navigation retrieval task.

We implemented an information retrieval system that indexes only site anchor text instead of the full-text of documents. Navigation retrieval for site anchor text has the following two advantages:
(1) The index size is very small.
(2) A user can retrieve uncrawled documents as well as crawled documents since the system needs only reference information.

Since it is necessary to maximize limited information to determine the relevancy of documents, in this paper we propose two relevancy measures: reference consistency and specificity of word combination.

2 Retrieval method

In this section we describe an extraction method for site anchor text and a determination method for the representativeness of Web pages. We also explain two measures for determining the relevancy of documents queried in the retrieval process.

2.1 Site anchor text and the representativeness of a Web page

Site anchor text is defined as the text in a link indicating the top page of a given Web site from external Web sites [1]. For instance, in Figure 1, page *a* in a domain named "www.d.com" is indicated from page *b* - *g*. To avoid ambiguity, we define "external Web sites" simply as sites whose domain name is different from the target page. Let Anchor(*a*,*b*) be the anchor text of link pointing from page *a* to *b*. Then the site anchor text to the page *a* is Anchor(*d*,*a*), Anchor(*e*,*a*), Anchor(*f*,*a*) and Anchor(*g*,*a*). Note that Anchor(*b*,*a*) and Anchor(*c*,*a*) are excluded from the site anchor text of page *a* because page *b* and *c* are in the same domain as page *a*.

We can also determine the representativeness of a given page *p* from the link structure shown in Figure 1. The representativeness of Rep(*p*) is determined as follows:

$$\mathrm{Rep}(p) = C \times T ,$$

where *C* is a citation frequency from an external Web

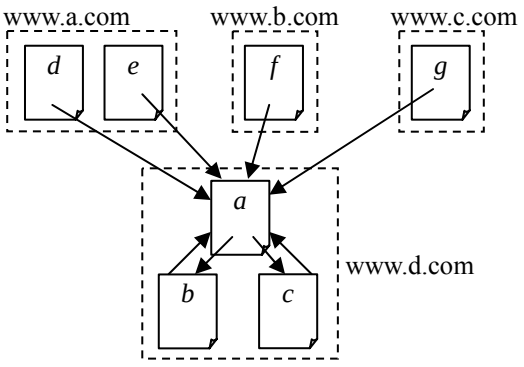

Figure 1. Example of site anchor text

site that indicates how many people accept the value of page p as a destination outside of their site. In Figure 1, the citation frequency C of page a is 4 because there are four links incoming from external domains: "www.a.com", "www.b.com" and "www.c.com".

T is the likelihood that the top page will be determined by applying heuristics to a given page. We used the following three simple heuristics:

(H1) Does the URL of the page consist only of the domain name?

(H2) Does the file name of the URL contain such a string as "index" or "default"?

(H3) Does the URL end with a slash "/"?

We calculated T as the weighted linear sum of these heuristics:

$$T = w_1 \times \delta_1 + w_2 \times \delta_2 + w_3 \times \delta_3 + w_4 ,$$

where w_1 to w_3 are the weight of each heuristic and w_4 is constant. In this paper we set w_1=1000, w_2=100, w_3=10, w_4=1 as a trial. And δ_m is 1 if the m-th heuristic is true or 0 if the m-th heuristic is false.

2.2 Retrieval process

The retrieval process for site anchor text follows:

Step 1: Parse the query designated by TopicPart (<TITLE> or <DESC>) with a Japanese morphological analysis system ChaSen [2]. We searched for nouns and unknown words.

Step 2: Determine the relevancy of page p to query q as Rel(p,q), and then calculate Score(p) as follows:

$$\text{Score}(p) = \text{Rep}(p) \times \text{Rel}(p, q) .$$

Step 3: Sort pages by Score(p).

To determine relevancy Rel(p,q), we used two measures: reference consistency and specificity of word combination. In next section, we will describe in more details these measures.

2.2.1 Reference consistency

Reference consistency is a measure that indicates the consistency of the top page of a given site recommended by an external Web site. For example, if a given site cover a myriad of topics, it can be assumed that the site anchor text for the top page of the site contains various words because these words are not so consistent. To illustrate, consider the links indicating the top page of a news site. The site anchor text may be headlines of news articles, titles or the URL of the site.

The news article on the top page is usually updated every day. Site anchor text using the headline seems to be inconsistent because it can be changed in response to a change in the article; otherwise it contains a lot of keywords covering various old news topics. Thus, the top page can be less relevant, even if the query matches the keywords in the site anchor text that is using news headlines. However, site anchor text using titles or URLs of the news site is consistent, so the relevancy of the top page can be improved if the query matches the keywords in the site anchor text using title or URL of the site.

We defined relevancy Rel(p,q) based on the reference consistency as follows:

$$\text{Rel}(p, q) = \sum_{t \in q} kw_t \times \left(\frac{f_t^{\,2}}{N_{sa}} \right),$$

where f_t is the frequency of word t in the site anchor text for page p, N_{sa} is the amount of site anchor text for page p, and kw_t is the weight of word t in query q. The word frequency rate in the site anchor text may be calculated by (f_t / N_{sa}); however, it treats pages in which word t appears 50 times in 100 site anchor texts and the page in which word t appears only once in 2 site anchor texts in the same manner. So we used (f_t^2 / N_{sa}) to give the former page higher relevancy.

Because word t is set in descending order of importance, we decided the weight kw_i of the i-th word t_i in query q as follows:

$$kw_i = 2^{(n_q - i)} ,$$

where n_q is the number of words in query q.

Google's PageRank [3] and Kleinberg's HITS [4] are two link-structure analysis algorithms. Our method has two main advantages: (1) it requires less calculation cost because reference consistency is obtained without any matrix-vector multiplication; and (2) it reduces topic drift, a common problem among link-structure analysis algorithms because tightly linked irrelevant pages tend to dominate relatively sparse relevant pages.

2.2.2 Specificity of word combination

The total size of the site anchor text index is much smaller than the original document set. Therefore, all of the words in a query rarely appear in the site anchor text. Since a query contains a lot of words when TopicPart is <DESC>, it is necessary to determine which word is important for the search. In traditional retrieval systems, Term Frequency Inverse Document Frequency (TF-IDF) is a typical weighting scheme based on frequency and specificity of the word [5].

However, TF-IDF tends to be biased toward words with high specificity if the frequency of each word is at the same level. Especially in site anchor text, each word appears only several times because the length of the site anchor text is very short. Figure 2 gives more details with an example of word combinations. Figure 2 describes groups of documents that include words t_1-t_3 in a given query q. Word t_1 has low specificity because it appears in a lot of documents. In contrast t_2 and t_3 have comparatively high specificity.

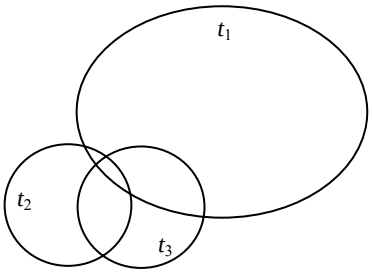

Figure 2. Example of word combination

Let D(t_1, t_2, t_3) be a document group that contains words t_1, t_2, and t_3. If the term frequency of each word is the same, D(t_1, t_2, t_3) will be the most highly scored for the query q=(t_1, t_2, t_3). Then, in Figure 2, D(t_2, t_3) tends to be the second highly scored document group based on TF-IDF because the IDF of t_2 and t_3 are high. However, D(t_1, t_2) is more highly specified than D(t_2, t_3), meaning that document groups which contain combinations of high IDF words are not always so specific.

Thus, we defined document relevancy Rel(p,q) based on the specificity of word combinations as follows:

$$\mathrm{Rel}(p,q) = \log \frac{N}{|\mathrm{D}(\tau \in p,q)|},$$

where N is the number of all pages, $|\mathrm{D}(\tau \in p,q)|$ is the number of pages that contains word group τ included in both page p and query q.

Relevancy based on reference consistency differs from relevancy based on specificity of word combinations. We used each measure in different systems independently instead of combining them into one system.

3 Evaluation

We compared the following four retrieval systems to estimate the efficiency of site anchor text and the two kinds of measures of document relevancy:

(1) A baseline system that indexes the complete text of crawled documents and determines relevancy by OKAPI [6] (hereafter OKA).

(2) Another baseline system that indexes the complete text of crawled documents and determines relevancy by giving high weight to words in anchor text [1] (hereafter ANC)

(3) A retrieval system that indexes only site anchor text and determines relevancy based on reference consistency (hereafter SAR).

(4) A retrieval system that indexes only site anchor text and determines relevancy based on specificity of word combination (hereafter SAS).

Note that two of the baseline systems (OKA and ANC) only return a list of crawled documents as a result of their search, while SAR and SAS return both crawled and uncrawled documents. Note also that since the OKA results were not submitted to the NTCIR formal run, they were not included in the pooling Web pages.

The experiment was conducted with a 100 GB document collection, 'NW100G-01', constructed on an NTCIR-3 WEB as the document set. The total size of the site anchor text was 94 MB, or one-thousandth of the original data.

The search result documents were subjectively judged “relevant,” “partially relevant” or “irrelevant,” and two evaluation scales, Discounted Cumulative Gain (DCG) and Weighted Reciprocal Rank (WRR), were calculated at the 10-document level based on relevancy judgment [7].

4 Results and discussion

The results of the DCG and the WRR of each system are shown in Figures 3 and 4. The system’s name is on the horizontal axis in Figures 3 and 4, and suffixes “TT” and “DS” indicate that <TITLE> or <DESC> was used for searching as TopicPart. Data series dcg.*x*-*y* in Figure 3 means that the DCG value was calculated by giving scores *x* and *y* for “relevant” and “partially relevant” pages respectively. Likewise, data series wrr.*x*-*y* in Figure 4 means that a WRR value was calculated by giving scores *x* and *y* for “relevant” and “partially relevant” pages respectively.

Figures 3 and 4 reveal that both DCG and WRR showed almost the same tendency; there is also not a sharp contrast between DCG and WRR. Our findings are as follows:

(1) Relevancy based on reference consistency (SAR) achieved higher accuracy than specificity of word combination (SAS) when TopicPart is <TITLE>, but SAS was better than SAR when TopicPart is <DESC>. This indicates that choosing word combination based on specificity is efficient when TopicPart is <DESC>; however, it is not necessary to choose the word combination when TopicPart is <TITLE> because words in <TITLE> are selected and listed in order of importance in the search by humans.

In addition, comparing different TopicPart in the same system shows that the accuracy of <TITLE> is higher than <DESC>. This also means that some words may become ‘noise’ when TopicPart is <DESC>, but there are fewer such words in <TITLE> because they are selected well by humans.

(2) The retrieval systems based on site anchor text (SAR and SAS) outperformed the baseline system (OKA) that does not give any weight to anchor text. This indicates that site anchor text has great advantages for navigation retrieval tasks.

(3) Another baseline system, (ANC) which gives weight to the anchor text, outperformed the retrieval

systems that indexed only site anchor text (SAR and SAS). Some important information in anchor text was lost when site anchor text was extracted because, for SAR and SAS, we extracted simple links from the sites whose domain name is different from the target page. For example, if there are different sites in the same domain such as "http://abc.jp/~usr1/" and "http://abc.jp/~usr2/", then links from a page in the former site to a page in the latter site are not used for extracting site anchor text. We have to identify carefully the boundaries of Web sites.

(4) Despite a very small index, SAR and SAS can achieve up to 91% accuracy, compared with ANC. Table 1 shows the accuracy rates of DCG and WRR for SAR-TT, SAS-TT, and ANC-TT. In Table 1, the accuracy rates for SAR/ANC and SAS/ANC range from 74% to 91%, meaning that site anchor text retrieval can be close to the anchor text weighted full-text retrieval.

Especially accuracy ratio tends to be higher in data series that give a score only for the "relevant" pages (dcg.3-0 and wrr.1-0). These results show that site anchor text retrieval can pinpoint "relevant" pages. In contrast, as "partially relevant" scores (e.g. dcg.3-3, wrr.1-1) increase, accuracy ratios decrease. It appears that ANC tends to return more "partially relevant" pages than SAR and SAS. Actually, the number of "relevant" pages in ANC's search result was only 1.2 times the number of "relevant" pages in SAR's search result; however, the number of "partially relevant" pages in ANC's search result was more than 3.6 times the number of "partially relevant" pages in SAR's search result.

Figures 5 and 6 show the modified values of DCG and WRR derived from only crawled documents. These facts in Figures 3 and 4 can also be observed in Figures 5 and 6. The main difference between Figures 3 and 4 and Figures 5 and 6 is that the accuracy gap increased between SAR, SAS, and ANC. This suggests that some "relevant" pages in the search result of site anchor text retrieval (SAR, SAS) were ignored because they were uncrawled. Therefore, the accuracy of SAR and SAS decreased compared to ANC. Or to put it the other way around, some un-crawled pages are "relevant," and relevancy for the uncrawled pages can be partly determined based on only reference information. In fact, the rate of uncrawled pages in "relevant" pages returned by SAR and SAS was 30% and 31%, respectively.

5 Conclusion

In this paper we discussed an information retrieval system that indexes only site anchor text to verify the efficiency of reference information in a navigation retrieval task. We proposed two relevancy measures, reference consistency and specificity of word combination, to maximize limited information.

Site anchor text retrieval outperformed simple full-text retrieval and achieved up to 91% accuracy, compared with anchor text weighted full-text retrieval, despite using one-thousandth less information.

Since site anchor text retrieval can pinpoint highly relevant documents, it can be applied for downsizing retrieval systems. In addition, our relevancy calculation method can be applied to traditional retrieval systems to improve search accuracy. In future work we will address the problem of Web site boundaries and identify the optimum value of parameters used in various calculations.

References

[1] Kenji Tateishi, Hideki Kawai, Susumu Akamine, Katsushi Matsuda, and Toshikazu Fukushima, "Evaluation of Web Retrieval Method Using Anchor Text", *Proceedings of the 3rd NTCIR Workshop Meeting*, pp. 25-29, 2002.

[2] Yuji Matsumoto, Akira Kitauchi, Tatsuo Yamashita, Yoshitaka Hirano, Hiroshi Matsuda, Kazuma Takaoka, and Masayuki Asahara, Morphological Analysis System ChaSen version 2.2.2 Manual, *available from* http://chasen.aist-nara.ac.jp/stable/doc/chasen-2.2.2.pdf

[3] S. Brin and L.Page, The anatomy of a large-scale Web search engine, *Proceedings of 7th International WWW Conference*, pp.101-117, 1999.

[4] Jon M. Kleinberg: Authoritative sources in a hyperlinked environment, *Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms*, pp. 668 - 677, 1998.

[5] Gerard Salton and Christopher Buckley, Term-weighting approaches in automatic text retrieval, *Information Processing and Management,* Vol. 24 (5), pp. 513-523, 1988.

[6] S.E.Robertson, S. Walker, Okapi/Keenbow at TREC-8, *The Eighth Text Retrieval Conference*, 1999.

[7] Koji Eguchi, Keizo Oyama, Emi Ishida, Noriko Kando, and Kazuko Kuriyama, Overview of the Web Retrieval Task at the Third NTCIR Workshop," *Proceedings of the Third NTCIR Workshop on Research in Information Retrieval, Automatic Text Summarization and Question Answering*, 2003.

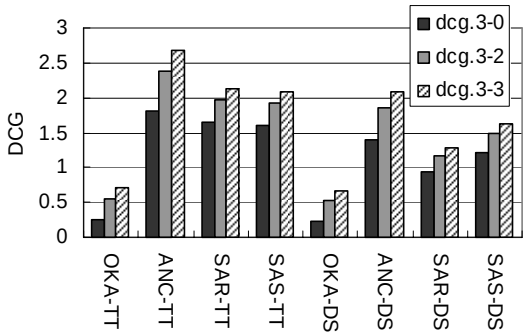

Figure 3. DCG value of each system

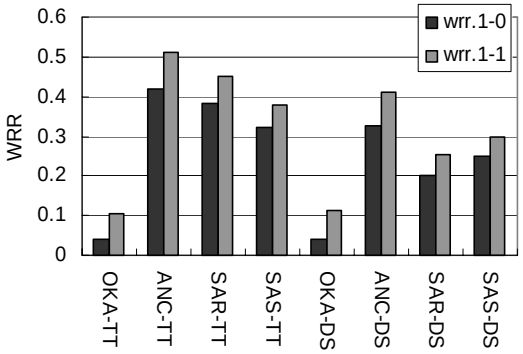

Figure 4. WRR value of each system

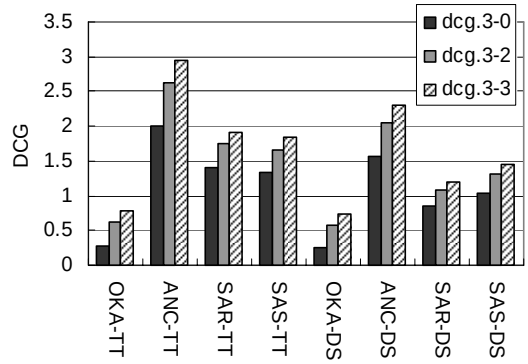

Figure 5. DCG for crawled pages

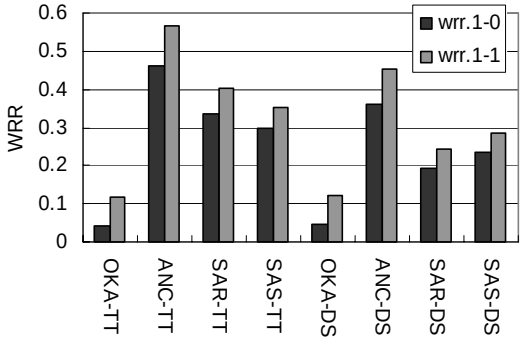

Figure 6. WRR for crawled pages

Table 1. Accuracy rate for each data series

	SAR/ANC	SAS/ANC
dcg.3-0	0.91	0.88
dcg.3-2	0.83	0.81
dcg.3-3	0.80	0.78
wrr.1-0	0.91	0.77
wrr.1-1	0.88	0.74