Verification of Effective Retrieval Method for Anchor Text on Navigational Retrieval

Kenji Tateishi Hideki Kawai Dai Kusui Toshikazu Fukushima[1]
NEC Internet Systems Research Labs.
8916-47, Takayama-cho, Ikoma-city, Nara, JAPAN
{k-tateishi@bq, h-kawai@ab, kusui@ct, t-fukushima@cj}.jp.nec.com

Abstract

We participated in NTCIR-5 WEB Navigational Retrieval Subtask(Navi-2) in order to verify the most effective retrieval method for the index of anchor texts by using a retrieval system that indexed only anchor texts instead of full texts of Web pages. We introduced retrieval methods that combine one or more of six retrieval measures: (a) anchor frequency (af), (b) reference consistency (rc), (c) query weight (qw), (d) page representativeness (rep), (e) site relevancy (sr), and (f) inverse anchor document frequency (iadf).

The experimental results revealed that: (1) it could be implied that the retrieval method that used only anchor frequency for the index of anchor texts was more effective than the retrieval method for the index of only full texts of Web pages, and that (2) the retrieval method that contained sr or iadf was effective for the index of anchor texts, and that sr was more effective than iadf.

Keyword: Anchor Text, Navigational Retrieval, Site Relevancy

1. Introduction

NTCIR-5 WEB Navigational Retrieval Subtask (Navi-2) 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 Web pages. Anchor text, text in a link to a Web page, summarizes the content of the Web page, and the number of anchor texts indicates the representativeness of the Web page.

We implemented a retrieval system that indexes only anchor texts instead of full texts of documents in NTCIR-4 [2], and verified the efficiency of anchor texts in the Navi-2. The result of experiment in NTCIR-4 shows the following:

- The retrieval system that indexes site anchor texts [1] achieves better performance than the one that indexes full texts of documents.
- The retrieval system that indexes site anchor texts [1] has a performance about 80% of the one that indexes both site anchor texts and full texts of documents. Site anchor texts enable the compact implementation of the retrieval system for Web pages.

We participated in Navi-2 in order to verify the effective retrieval method for the index of anchor texts. Specifically, we introduced retrieval methods that combine one or more of six retrieval measures: (a) anchor frequency (*af*), (b) reference consistency (*rc*), (c) query weight (*qw*), (d) page representativeness (*rep*), (e) site relevancy (*sr*), and (f) inverse anchor document frequency (*iadf*). The retrieval methods that combined four measures (a)-(d) were also used in the NTCIR-4 system, but the efficiency of each measure is not still clear. In this paper, we propose a new retrieval measure (e) in NTCIR-5, and use (f), a similar measure to idf, for comparison with (e).

2. Retrieval Method

Our system participating in NTCIR-5 uses retrieval methods that combine one or more of the six retrieval measures from Section 2.1 to 2.6. Comparisons between these methods will clarify which measure works well as a retrieval method for the index of anchor texts. Note that we also participated in work on the retrieval method for NTCIR-4 that combined the four measures described in Section 2.1 to 2.4.

2.1 Anchor Frequency (*af*)

Anchor frequency *af(p,t)* is a similar measure to term frequency, and is defined as the number of anchor texts that contain term *t* in links to a page *p*. Although *af* is one of the simplest measures, it is expected to be effective in that it is related to both relevance and representativeness of the Web page with respect to term *t*.

1 His present affiliation and address are the following:
NEC Laboratories, China.
11F, Building A, Innovation Plaza, Tsinghua Science Park, 1 Zhongguancun East Road, Beijing 100084, China.

2.2 Reference Consistency (rc)

Reference consistency is a measure that indicates the consistency of anchor tests in links from other Web pages to Web page *p* [2], and is defined as following formula:

$$rc(p,t) = \frac{af(p,t)}{N_a},$$

where N_a indicates the total number of anchor texts in links to Web page *p*.

Reference consistency is related to the relevance of the Web page *p* with respect ot the term. While anchor texts in links to a Web page are usually composed of many kinds of terms, the term used most commonly over anchor texts can express the most important meaning of the Web page *p*.

2.3 Query Weight (*qw*)

Query weight indicates the importance of term *t* in query terms. While *qw(t)* can be generally obtained by the frequency of term *t* in query terms, we selected the following formula because the importance of term *t* is already given in the NTCIR-5 search topics, where the order of query terms denotes it [2]:

$$qw(t) = 2^{(n_q - order(t))},$$

where n_q indicates the number of query terms, and *order(t)* represents the order of term *t* from the first query term.

2.4 Page Representativeness (*rep*)

Page representativeness indicates the representativeness of Web page *p*, and is defined as the following [2]:

$$rep(p) = C \times T,$$

where *C* is a citation frequency from external Web sites, indicating how many people recognize the value of Web page *p*, and *T* is the likelihood of Web page *p* to be a top page, obtained by using heuristics based on its URL. 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 liner sum of these heuristics:

$$T = w_1 x H1 + w_2 x H2 + w_3 x H3 + w_4,$$

where w_1 to w_3 are the weight of each heuristic and w_4 is constant. In this paper we set w_1=1,000, w_2=100, w_3=10, and w_4=1. Furthermore, H1, H2, and H3 are 1 if the heuristic is true or 0 if false.

2.5 Site Relevancy (*sr*)

Site relevancy is a measure that indicates the relevancy between term *t* and Web site *s* to which Web page *p* belongs, and is defined as the following formula:

$$sr(p,t) = \frac{2 \times df(t,s)}{2 \times df(t,s) + df(\neg t,s) + df(t,\neg s)},$$

where *df(t,s)* indicates the number of Web pages on the same domain as Web page *p* to which anchor texts in links contain the term t, *df(¬t,s)* denotes the number of Web pages on the same domain as Web page *p* to which anchor texts in links do not contain term *t*, and *df(t, ¬s)* indicates the number of Web pages on different domains from Web page *p* to which anchor texts in links contain term *t*. This formula is equivalent to calculating the similarity between term *t* and Web site *s* by the Dice coefficient.

Both *sr* and *iadf* in Section 2.6 share the same effect in that both can reduce the importance of common terms that appear in many Web sites. However, while *iadf* does not work in queries composed of just one term, *sr* gives different values even for one term query according to the relevancy between the term and the Web site. Therefore, *sr* is expected to be a better measure in Navi-2 where many one-term queries are included.

2.6 Inverse Anchor Document Frequency (*iadf*)

Inverse anchor document frequency (*iadf*), a similar measure to idf, indicates the generality of *t*, and is defined as the following:

$$iadf(t) = \log_2 \frac{N}{df(t)},$$

where *N* means the total number of Web pages, and *df(t)* indicates the number of Web pages to which anchor texts in links contain term *t*. When *iadf* is small, the term is regarded as a general term and is given low importance.

3. Experiment

We compare the performance between retrieval methods that combine one or more of retrieval measures from Section 2.1 to 2.6 in order to clarify which measure is most effective for the index of anchor texts.

3.1 Document Data Set

We indexed anchor texts from 1.36 TB of Web pages distributed by NTCIR-5[1]. We selected only anchor texts in links from different domains to a Web page[2]. The total number of anchor texts was 4,143,788. We used PostgreSQL for the indexing, a process that took about two weeks on a Linux PC with two Pentium II CPUs and 8 GB of memory.

3.2 Query

We used 268 Japanese search topics distributed by NTCIR-5 for a formal run. Each search topic is composed of three parts: Title, Description, and Narrative. We used only the Title part, each of which comprises a maximum of three keywords, with the importance of the keywords decreasing from left to right.

3.3 Retrieval Method

We applied the seven retrieval methods shown in Table 1, which were created by combining one or more of the measures in Section 2[3].

3.4 Evaluation Method

We used the Discounted Cumulative Gain (DCG) value and the Weighted Reciprocal Rank (WRR) value by [illegible]-ranked Web pages [3]. DCG and WRR differ in that DCG calculates the cumulative sum score according to both the rank and the relevancy grade by the [illegible]10 search results whereas WRR calculates the [illegible]ximum score. However, since it is common that both tend to be large values when highly relevant Web pages dominate the upper ranks of the top ten, the retrieval methods in Section 3.3 should show efficiencies in both evaluation methods.

DCG and WRR are allowed to set the weights for each relevancy grade: highly relevant, fairly relevant, and partially relevant. For example, dcg-3-3-0 gives these weights as (highly relevant, fairly relevant, partially relevant) = (3,3,0). In this paper, we used dcg-3-0-0, dcg-3-3-0, dcg-3-2-0, wrr-1-0-0, and wrr-1-1-0.

4. Experimental Results and Discussion

Table 1 shows the experimental results, and we analyze the efficiency of each retrieval measure in Section 4.1 to 4.5.

As mentioned in Section 3, the index we created contains errors, but we expect that the reliability of this experiment is maintained for the following reasons:

(1) There could remain many anchor texts in links to representative Web pages that Navi-2 targets because NTCIR-5 is supposed to index terabyte-sized Web pages.

(2) As we mention in Section 4.1 latter, RM6 in NTCIR-5 and the same retrieval method in NTCIR-4 resulted in approximately the same performance.

4.1 Efficiency of Anchor Frequency (*af*)

On NTCIR-4, we evaluated a retrieval system that indexes only full-text Web pages [2], resulting in a DCG value of about 1.1 at most and a WRR of about 0.15 at most; K1300-12 on NTCIR-5 surpasses this performance. On the other hand, RM6 on NTCIR-5 was also applied to the system on NTCIR-4, giving a DCG value of at most approximately 2.1, and a WRR of at most about 0.45, similar to the RM6 result. From this result, we infer that NTCIR-4 and NTCIR-5 have similar DCG and WRR values on the same index and on the same retrieval method. Therefore, we infer that the retrieval method using only *af* for the index of anchor texts is more effective than the retrieval method for only full text of Web pages.

4.2 Efficiency of Reference Consistency (*rc*)

A comparison between RM1 and RM2 reveals that the performance of RM2, a combination of *af* and *rc*, is lower than RM1, the retrieval method using only *af*, on both DCG and WRR. Therefore, the efficiency of *rc* could not be observed on NTCIR-5.

4.3 Efficiency of Query Weight (*qw*)

A comparison between RM2 and RM3 shows that

1 The index we created includes the following two errors:

(1) Only one of ten anchor texts could be extracted from distributed Web pages.

(2) The same anchor text from the same domain to a Web page was counted only once. For example, when there are two anchor texts labeled “NEC” from www.a.com to Web page *p*, three from www.b.com, and two from www.c.com, anchor text labeled “NEC” in Web page *p* should be counted as seven . However, it is incorrectly counted as three.

2 We have prepared another index including anchor texts from the same domain in addition to from different domains, but we stopped using it because of the lower reliability of the index caused by error (2) above.

3 In fact, we submitted other runs for the formal run, but we have omitted seven runs in this paper:.. These runs searched the index, including anchor texts, from the same domain, and we judged these runs as being less reliable due to error (2) above.

Table 1. Retrieval methods applied to the formal run and their results.

ID	Retrieval Method	DCG. 3-0-0	DCG. 3-2-0	DCG. 3-3-0	WRR. 1-0-0	WRR. 1-1-0
RM1	$\sum_t af$	1.583	1.918	2.086	0.3933	0.4728
RM2	$\sum_t (af \times rc)$	1.492	1.806	1.964	0.3654	0.4434
RM3	$\sum_t (af \times rc \times qw)$	1.490	1.801	1.957	0.3700	0.4500
RM4	$\sum_t (af \times rc \times qw \times iadf)$	1.497	1.812	1.969	0.3708	0.4515
RM5	$\sum_t (af \times rc \times qw \times sr)$	1.607	1.946	2.115	**0.3962**	**0.4813**
RM6	$rep \times \sum_t (af \times rc \times qw)$	1.513	1.845	2.011	0.3626	0.4503
RM7	$rep \times \sum_t (af \times rc \times qw \times sr)$	**1.616**	**1.965**	**2.139**	0.3881	0.4744

the performance of RM3, a combination of *af*, *rc*, and *qw*, is higher than that of RM2, which is a combination of *af* and *rc*, on WRR, but is lower on DCG. Therefore, the efficiency of *qw* could not be observed consistently on NTCIR-5.

4.4 Efficiency of Site Relevancy (*sr*) and Inverse Anchor Document Frequency (*iadf*)

A comparison between RM3 and RM5 and RM4 reveals that the performance of RM5, a combination of *af*, *rc*, *qw*, and *sr*, and RM4, which is a combination of *af*, *rc*, *qw*, *iadf*, are higher than that of RM3, a combination of *af*, *pc*, *qw*, on both DCG and WRR. Moreover, the improvement of RM5 is greater than that of RM4. Therefore, the retrieval method containing *sr* or *iadf* was effective for the index of anchor texts, with *sr* being more effective than *iadf*.

4.5 Efficiency of Page Representativeness (*rep*)

A comparison between RM3 and RM6 shows that the performance of RM6, a combination of *af*, *rc*, *qw*, and *rep*, is higher than that of RM3, which is a combination of *af*, *rc*, and *qw*, on DCG, but lower on WRR. Therefore, the efficiency of *rep* could not be observed consistently on NTCIR-5.

5. Conclusion

We participated in Navi-2 in order to verify the most effective retrieval method for the index of anchor texts by using a retrieval system that indexed only anchor texts instead of the full text of Web pages. We introduced retrieval methods that combine one or more of six retrieval measures: (a) anchor frequency (*af*), (b) reference consistency (*rc*), (c) query weight (*qw*), (d) page representativeness (*rep*), (e) site relevancy (*sr*), and (f) inverse anchor document frequency (*iadf*). The experimental results revealed the following:

- It can be inferred that the retrieval method using only *af* for the index of anchor texts is more effective than the retrieval method for only the full text of Web pages.
- The retrial method including *sr* or *iadf* is effective for the index of anchor texts, and *sr* is more effective than *iadf*.
- The efficiency of *rc* for the index of anchor texts cannot be observed on NTCIR-5.
- The efficiency of *rep* for the index of anchor texts cannot be observed consistently on NTCIR-5 because WRR and DCG show different results.

Site relevancy, the efficiency of which can be observed on NTCIR-5, distinguishes Web sites using these domains. However, there are many domains that contain different Web sites, and there are many Web sites that contain different domains. Moreover, it is expected that site relevance will be effective even for a retrieval system that indexes the full text of Web pages. In future work, we plan to devise an algorithm for dividing Web sites more intelligently, and to verify the efficiency of site relevancy with respect to the retrieval system indexing the full text of Web pages.

References

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

[2] Hideki Kawai, Kenji Tateishi, Toshikazu Fukushima, “Navigation Retrieval with Site Anchor Text”, In Proceedings of the 4th NTCIR Workshop Meeting, 2004.

[3] 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.