Exploiting Anchor Text for the Navigational Web Retrieval at NTCIR-5

Atsushi Fujii † Katunobu Itou ‡ Tomoyosi Akiba * Tetsuya Ishikawa †
† Graduate School of Library, Information and Media Studies, University of Tsukuba
‡ Graduate School of Information Science, Nagoya University
* Department of Information and Computer Sciences, Toyohashi University of Technology
fujii@slis.tsukuba.ac.jp

Abstract

In the Navigational Retrieval Subtask 2 (Navi-2) at the NTCIR-5 WEB Task, a hypothetical user knows a specific item (e.g., a product, company, and person) and requires to find one or more representative Web pages related to the item. This paper describes our system participated in the Navi-2 subtask and reports the evaluation results of our system. Our system uses three types of information obtained from the NTCIR-5 Web collection: page content, anchor text, and link structure. Specifically, we exploit anchor text in two perspectives. First, we compare the effectiveness of two different methods to model anchor text. Second, we use anchor text to extract synonyms for query expansion purposes. We show the effectiveness of our system experimentally.

Keywords: *Navigational Web retrieval, Anchor text model, Link structure analysis, NTCIR*

1 Introduction

In the Navigational Retrieval Subtask 2 (Navi-2) at the NTCIR-5 WEB Task, a hypothetical user knows a specific item (e.g., a product, company, and person) and requires to find one or more representative Web pages related to the item [10]. This subtask is fundamentally the same as the Navigational Retrieval Subtask 1 (Navi-1) at NTCIR-4 [9]. However, the numbers of topics and documents were independently increased at NTCIR-5. The organizers provided the participants with 400 topics and a document collection consisting of approximately one hundred million pages.

This paper describes our system participated in the Navi-2 subtask and reports the evaluation results of our system. Our system uses three types of information obtained from the NTCIR-5 Web collection: page content, anchor text, and link structure. Specifically, we exploit anchor text in two perspectives. First, we compare the effectiveness of two different methods to model anchor text. Second, we use anchor text to extract synonyms for query expansion purposes.

2 System Description

2.1 Overview

In the TREC Web Track, a combination of page content, anchor text, and link structure was arguably effective for the home/named page finding task. In the NTCIR-4 WEB task, a combination of page content and anchor text was effective for the Navi-1 subtask. Thus, as with existing methods mainly targeting these tasks [2, 8, 13, 14], we use content, anchor, and link structure information.

However, we do not model these three types of information in a single framework. Instead, these information types are used independently to produce three ranked document lists, in each of which documents are sorted according to the score with respect to a query. These lists are integrated into a single list and up to the top N documents are used in the final retrieval result. In the formal run of the Navi-2 subtask, $N = 100$.

However, because the scores computed by the three types of information have different interpretations and ranges, it is difficult to combine these scores in a mathematically founded method. Thus, we use an ad-hoc method and re-rank each document by a weighted harmonic mean of the ranks in the three lists. We compute the final score for document d, $S(d)$, as in Equation (1).

$$S(d) = \frac{1}{\lambda_c \frac{1}{R_c(d)} + \lambda_a \frac{1}{R_a(d)} + \lambda_s \frac{1}{R_s(d)}} \qquad (1)$$
$$\lambda_c + \lambda_a + \lambda_s = 1,\ \lambda_c \geq 0,\ \lambda_a \geq 0,\ \lambda_s \geq 0$$

$R_c(d)$, $R_a(d)$, and $R_s(d)$ are the ranks of d in the content-based, anchor-based, and structure-based lists, respectively. λ_c, λ_a, and λ_s, which range from 0 to 1, are parametric constants to control the effects of $R_c(d)$, $R_a(d)$, and $R_s(d)$ in producing the final list, respectively.

In Sections 2.2–2.4, we explain the retrieval methods using the three information types, respectively.

2.2 Content-based Retrieval

To use page content for retrieval purposes, we index the documents in the Web collection by words and bi-words. We use ChaSen[1] to perform morphological analysis on the document files from which HTML tags were removed by the organizers and extract nouns, verbs, adjectives, out-of-dictionary words, and symbols as index terms. We use Okapi BM25 [11] to compute the content-based score for each document with respect to a query, as in Equation (2).

$$\sum_{t \in q} f_{t,q} \cdot \frac{(K+1) \cdot f_{t,d}}{K \cdot \{(1-b) + \frac{dl_d}{b \cdot avgdl}\} + f_{t,d}} \cdot \log \frac{N - n_t + 0.5}{n_t + 0.5} \tag{2}$$

$f_{t,d}$ and $f_{t,d}$ denote the frequency with which term t appears in query q and document d, respectively. N and n_t denote the total number of documents in the Web collection and the number of documents containing term t, respectively. dl_d denotes the length of d, and $avgdl$ denotes the average length of documents in the collection. We set $K = 2.0$ and $b = 0.8$, respectively, as these values were used in the literature [6].

2.3 Anchor-based Retrieval

To use anchor text for retrieval purposes, we index the anchor text in the Web collection by words and compute the score for each document with respect to a query. We compute the probability that document d is the representative page for the item expressed by query q, $P(d|q)$. We elaborate on the computation of $P(d|q)$ in Section 3.

2.4 Structure-based Retrieval

To use link structure for retrieval purposes, we analyze the structure of links in the Web collection and compute the score for each document. We use PageRank [1] to compute the probability that a user surfing on the Web visits document d, $P(d)$. Unlike the content-based and anchor-based scores, the structure-based score is independent of the query. Thus, we use the content-based and anchor-based scores to collect candidate documents and sort only these documents according to the value of $P(d)$.

For link structure analysis, we use the "linklist" files provided by the organizers. However, because the computation of PageRank is prohibitive, we discarded documents for which either of the number of inlinks or the number of outlinks is below m. We experimentally set $m = 5$ with no particular reason. As a result, approximately 29M documents were used for the computation of PageRank. For the remaining documents, $P(d) = 0$.

[1] http://chasen.aist-nara.ac.jp/index.html.en

3 Exploiting Anchor Text

3.1 Overview

To utilize anchor text in our system, we compute the probability that document d is the representative page for the item expressed by query q, $P(d|q)$. Given q, the task is to select the d that maximizes $P(d|q)$, which is transformed as in Equation (3) using Bayes' theorem.

$$\arg\max_d P(d|q) = \arg\max_d P(q|d) \cdot P(d) \tag{3}$$

We estimate $P(d)$ as the probability that d is retrieved by an anchor text randomly selected from the Web collection. $P(d)$ is calculated as the ratio of the number of links to d in the Web collection and the total number of links in the Web collection.

We assume the independence of the terms in q and approximate $P(q|d)$ as in Equation (4).

$$P(q|d) = \prod_{t \in q} P(t|d) \tag{4}$$

To extract term t in q, we use ChaSen and extract index terms as in the content-based indexing (see Section 2). However, unlike the content-based indexing, we use only words as t. We elaborate on two alternative models to compute $P(t|d)$ in Section 3.2.

We extracted anchor text from documents in the Web collection. However, because pages in the same Web server are often maintained by the same person or the same group of people, links and anchor texts between those pages can potentially be manipulated so that their pages can be retrieved in response to various queries. To resolve this problem, we discarded the anchor text used to link pages in the same server. Because we used a string matching method to identify servers, variants of the name of a single server, such as alias names, were considered different. Additionally, even if a page links to another page more than once, we extracted only the first anchor text.

Because each anchor text is usually shorter than a document, the mismatch between a term in an anchor text and a term in a query potentially decreases the recall of the anchor-based retrieval. A query expansion method is effective to resolve this problem.

However, in the Navi-2 subtask the precision is more important than the recall. In view of the above discussion, we expand a query term only if $P(t|d)$ is not modeled in our system. In such a case, we use a synonym of t, s, as a substitution of t and approximate $P(t|d)$ as in Equation (5).

$$\begin{aligned} P(t|d) &= P(t|s,d) \cdot P(s|d) \\ &\approx P(t|s) \cdot P(s|d) \end{aligned} \tag{5}$$

$P(t|s)$ denotes the probability that s is replaced with t. To derive the second line of Equation (5), we assume that the probability of s being replaced with t is

independent of d. The interpretation and computation of $P(s|d)$ are the same as $P(t|d)$, which is explained in Section 3.2. We elaborate on the methods to extract synonyms and to compute $P(t|s)$ in Section 3.3.

However, if no synonyms of t are modeled in our system, we need a different smoothing method; otherwise the product calculated by Equation (4) becomes zero. For smoothing purposes, we replace $P(t|d)$ with $P(t)$, which is the probability that a term randomly selected from the Web collection is t. Thus, if mismatched query terms are general words that frequently appear in the collection, such as "system" and "page", the decrease of $P(q|d)$ in Equation (4) is small. However, if mismatched query terms are low frequent words, which are usually effective for retrieval purposes, $P(q|d)$ decreases significantly.

3.2 Modeling Anchor Text

To compute $P(t|d)$ in Equation (4), we use two alternative models.

In the first model, a set of all anchor texts linking to d, $\mathbf{A}_d$, is used as a single document, D, which is used as surrogate content of d. $P(t|d)$ is computed as the ratio of the frequency of t in D to the total frequency of all terms in D [14].

In the second model, which is proposed in this paper, each anchor text $a \in \mathbf{A}_d$ is used independently and $P(t|d)$ is computed as in Equation (6).

$$P(t|d) = \sum_{a \in \mathbf{A}_d} P(t|a) \cdot P(a|d) \qquad (6)$$

$P(t|a)$ denotes the probability that a term randomly selected from $a \in \mathbf{A}_d$ is t. We compute $P(t|a)$ as the ratio of the frequency of t in a to the total frequency of all terms in a. $P(a|d)$ denotes the probability that an anchor text randomly selected from $\mathbf{A}_d$ is a. We compute $P(a|d)$ as the ratio of the frequency with which a links to d and the total frequency of all anchor texts in $\mathbf{A}_d$. To improve the efficiency of the computation for Equation (6), we consider only such a that includes t.

We call the first and second models "document model" and "anchor model", respectively.

We illustrate the difference of these two models comparing the following two cases. In the first case, d is linked from four anchor texts a_1, a_2, a_3, and a_4. Each a_i consists of a single term t_i. In the second case, d is linked from two anchor texts a_1 and a_2. While a_1 consists of t_1, t_2, and t_3, a_2 consists of t_4.

In the document model, $P(t_i|d)$ is $\frac{1}{4}$ for each t_i in either case. However, this calculation is counterintuitive. While in the first case each t_i is equally important, in the second case t_4 should be more important than the other terms, because t_4 is equally informative as a set of t_1, t_2, and t_3. In the anchor model, while $P(t_4|a_2)$ is 1, $P(t_i|a_1)$ $(i = 1, 2, 3)$ is $\frac{1}{3}$ for the second case. Thus, if $P(a_1|d)$ and $P(a_2|d)$ are equal, $P(t_4|d)$ becomes greater than $P(t_i|d)$ $(i = 1, 2, 3)$.

We further illustrate the difference of these two models using a hypothetical example. We use "http://www.yahoo.co.jp/" as d and we assume that d is linked from the following three anchor texts: a_1 = {Yahoo, Japan}, a_2 = {*yafuu*}, and a_3 = {Yahoo}. Here, "*yafuu*" is a romanized Japanese translation corresponding to "Yahoo". We also assume that the probability of $P(a_i|d)$ is uniform and thus $P(a_i|d) = \frac{1}{3}$ for any a_i.

In the document model, $P(t|d)$ for each term is as follows:

- $P(\text{Yahoo}|d) = \frac{1}{2}$,
- $P(yafuu|d) = \frac{1}{4}$,
- $P(\text{Japan}|d) = \frac{1}{4}$.

In the anchor model, $P(t|d)$ for each term is calculated as follows:

- $P(\text{Yahoo}|d) = 1 \times \frac{1}{3} + \frac{1}{2} \times \frac{1}{3} = \frac{1}{2}$,
- $P(yafuu|d) = 1 \times \frac{1}{3} = \frac{1}{3}$,
- $P(\text{Japan}|d) = \frac{1}{2} \times \frac{1}{3} = \frac{1}{6}$.

Unlike the document model, in the anchor model $P(yafuu|d)$ is greater than $P(\text{Japan}|d)$. In real world, "*yafuu*" is more effective than "Japan" in searching for "http://www.yahoo.co.jp".

In summary, the anchor model is more intuitive than the document model. We compare the effectiveness of these two models quantitatively in Section 4.

3.3 Extracting Synonyms

When more than one anchor text link to the same Web page, these texts generally represent the same or similar content. For example, "google search" and "*guuguru kensaku*" (romanized Japanese translation corresponding to "google search") can independently be used as an anchor text to produce a link to "http://www.google.co.jp".

While existing methods to extract translations use documents as a bilingual corpus [12], we use a set of anchor texts linking to the same page as a bilingual corpus. Because anchor texts are short, the search space is limited and thus the accuracy is possibly higher than that for general translation extraction tasks. In principle, both translations and synonyms can be extracted by our method. However, in practice we target only transliteration equivalents, which can usually be extracted with a high accuracy relying on phonetic similarity. We target words in European languages (mostly English) and their translations spelled out with Japanese *Katakana* characters.

Our method consists of the following three steps:

1. identification of candidate word pairs,
2. extraction of transliteration equivalents,
3. computation of $P(t|s)$ used in Equation (5).

In the first step, we identify words written with the Roman alphabet or the *Katakana* alphabet. These words can systematically be identified in the EUC-JP character code.

In the second step, for any pairs of European word e and Japanese *Katakana* word j, we examine whether or not j is a transliteration of e. For this purpose, we use our transliteration method [4, 5], which can process any of Japanese, English, and Korean as both the source and target languages.

If either of e or j can be transliterated into its counterpart by our method, we extract "(e,j)" as a transliteration equivalent pair. We compute the probability that s is a transliteration of t, $p(t|s)$, and select the t that maximizes $p(t|s)$, which is transformed as in Equation (7) using Bayes' theorem.

$$\arg\max_t p(t|s) = \arg\max_d p(s|t) \cdot p(t) \qquad (7)$$

$p(s|t)$ denotes the probability that t is transformed into s on a phone-by-phone basis. If $p(s|t) = 0$, t is not a transliteration of s. $p(t)$, which denotes the probability that t is generated as a word in the target language, is modeled by a word unigram produced from the anchor text. $p(t)$ is determined by Equation (8).

$$p(t) = \begin{cases} 1 & \text{if } t \text{ is the counterpart of } s \\ 0 & \text{otherwise} \end{cases} \qquad (8)$$

In summary, we extract "(e,j)" as a transliteration equivalent pair, only if $p(e|j)$ or $p(j|e)$ becomes a positive value. Because the transliteration is not an invertible operation, we compute both $p(e|j)$ and $p(j|e)$ to increase the recall of the synonym extraction.

We do not use $p(t|s)$ as $P(t|s)$ in Equation (5), because we need the probability that t can be a substitution for s when used in an anchor text. Equation (7) is used only for extracting transliteration equivalents. Thus, in the final step, we compute $P(t|s)$ as in Equation (9).

$$P(t|s) = \frac{F(t,s)}{\sum_{r \neq s} F(r,s)} \qquad (9)$$

$F(t,s)$ denotes the frequency that t and s independently appear in different anchor texts linking to the same document. For transliteration equivalent "(e,j)", we compute both $P(e|j)$ and $P(j|e)$.

4 Evaluation

4.1 Evaluation Method

As performed in the formal run of the Navi-2 subtask, we used DCG (Discounted Cumulative Gain) [7] and WRR (Weighted Reciprocal Rank) [3] as evaluation measures and investigated the effectiveness of each component in our system. We fixed several bugs of our system after the formal run and consequently experimental results were marginally improved. In this paper, we report only the newest results.

For each topic, we used only the terms in the "TITLE" field as a query.

In the relevance judgment performed by the organizers, relevance of each document with respect to a topic was judged by "relevant (A)", "partially relevant (B)", or "irrelevant". Search topics are classified as to which types of relevant documents were found during the relevance judgment process. While in "TYPE=A" at least one relevant document was found, in "TYPE=AB" at least one relevant or partially relevant document was found. Thus, by definition each topic can be classified into one or more types. The numbers of topics for "TYPE=A" and "TYPE=AB" were 269 and 308, respectively.

To calculate the DCG and WRR for each method, we used the official evaluation tool provided by the organizers. For the parametric constants in this tool, we used the default values set by the organizers. The cutoff rank was 10. To calculate the DCG and WRR, the parameters (or scores) for relevant and partially relevant documents, "(X,Y)", must be specified. While for DCG we used (3,0) and (3,2) independently, for WRR we used (1,0) and (1,1) independently.

4.2 Results

Table 1 shows the DCG and WRR for different combinations of components in our system. In Table 1, "DCG-X-Y" and "WRR-X-Y" denote the DCG and WRR calculated using parameter set "(X,Y)". Each method is represented by one or more components denoted as follows:

- AM: the anchor model in the anchor-based retrieval (Section 3.2),
- DM: the document model in the anchor-based retrieval (Section 3.2),
- Syn: the query expansion using synonyms (Section 3.3),
- C: the content-based retrieval (Section 2.2).

By comparing the document and anchor models, AM outperformed DM and AM+Syn outperformed DM+Syn except for WRR-1-1. Thus, the anchor model was usually effective than the document model disregarding the use of the synonym-based query expansion.

By comparing AM and AM+Syn (or DM and DM+Syn), the synonym-based query expansion was

Table 1. Evaluation results for different methods.

	TYPE=A				TYPE=AB			
Method	DCG-3-0	DCG-3-2	WRR-1-0	WRR-1-1	DCG-3-0	DCG-3-2	WRR-1-0	WRR-1-1
AM+Syn+C	2.522	2.979	0.605	0.661	2.203	2.674	0.529	0.602
AM+Syn	2.499	2.925	0.600	0.657	2.182	2.619	0.524	0.597
AM	2.464	2.885	0.596	0.650	2.152	2.584	0.521	0.591
DM+Syn	2.460	2.881	0.593	0.654	2.148	2.580	0.518	0.598
DM	2.431	2.847	0.590	0.650	2.124	2.551	0.516	0.594
C	0.381	0.665	0.080	0.116	0.333	0.645	0.070	0.113

marginally improved the DCG and WRR of the anchor-based retrieval.

By comparing the variations of the anchor-based retrieval (i.e., DM, DM+Syn, AM, and AM+Syn), AM+Syn was most effective in terms of the DCG and WRR.

By comparing the content-based retrieval and the anchor-based retrieval, the DCG and WRR of C were generally well below those of the remaining methods. Thus, in the navigational Web retrieval the anchor-based retrieval was effective than the content-based retrieval. However, when we combined the both retrieval methods in AM+Syn+C, the DCG and WRR of AM+Syn were generally improved.

In AM+Syn+C, we set $\lambda_c = 0.2$, $\lambda_a = 0.8$, and $\lambda_s = 0$ for Equation (1), which were the optimal values determined through preliminary experiments. In other words, the structure-based retrieval was not effective in our experiments. We observed that the effectiveness of the anchor-based score was significant and thus the structure-based score, which is independent of the query, generally decreased the DCG and WRR.

In summary, a) the anchor text model, b) the query expansion using automatically extracted synonyms, and c) a combination of the anchor-based and content-based retrieval methods were independently effective to improve the accuracy of the navigational Web retrieval task. Although the improvement of each enhancement was small, when used together the improvement was noticeable.

4.3 Topic-by-topic Analysis

We further investigate the effectiveness of each method evaluated in Section 4.2 on a topic-by-topic basis. In Table 2, the values of "X / Y" in the DCG and WRR columns denote the number of topics improved by the methods in the "Methods" column.

By comparing DM+Syn and AM+Syn, the improvement by AM+Syn was observed for more topics than DM+Syn except for WRR-1-1.

By comparing AM and AM+Syn, the DCG and WRR were varied for a small number of topics. For these topics, we describe the topic ID and the terms expanded in AM+Syn. Here, we romanize Japanese *Katakana* words.

- AM > AM+Syn

  1041: UNESCO → *yunesuko*

- AM < AM+Syn

  1097: *ekisaito* → excite

  1131: *dansu* → dance, *diraito* → delight

  1138: *toyota* → toyota, *chiimu* → team

  1172: *direkutori* → directory

Although all the above transliterations are correct, for Topic 1041 the query expansion decreased the DCG of AM. While for Topics 1097 and 1131 AM did not retrieve relevant documents in the top ten, the query expansion successfully retrieved relevant documents for these topics.

By comparing AM+Syn and AM+Syn+C, the improvement by AM+Syn was usually observed for more topics than AM+Syn+C, although as in Table 1 AM+Syn+C outperformed AM+Syn in the total DCG. By combining the content-based retrieval with AM, the number of topics for which a relevant document was retrieved in the top ten documents was increased. In other words, the content-based retrieval improved the DCG and WRR for a small number of topics, but the improvement for each topic was great.

By comparing C and AM+Syn+C, we reconfirmed that in the navigational Web retrieval, the anchor-based retrieval was more effective than the content-based retrieval.

4.4 Analysis by Topic Subcategories

In the Navi-2 subtask, the topics were categorized by the organizers from the following three perspectives.

- Type: complexity of representing the information need as a query

  1: single keyword or single phrase, 2: combination of keywords, 3: incomplete representation

Table 2. Topic-by-topic comparison.

Methods	TYPE=A				TYPE=AB			
	DCG-3-0	DCG-3-2	WRR-1-0	WRR-1-1	DCG-3-0	DCG-3-2	WRR-1-0	WRR-1-1
DM+Syn / AM+Syn	15 / 23	20 / 31	11 / 13	12 / 8	15 / 23	21 / 33	11 / 13	14 / 10
AM / AM+Syn	1 / 4	1 / 4	0 / 3	0 / 3	1 / 4	1 / 4	0 / 3	0 / 3
AM+Syn / AM+Syn+C	29 / 13	46 / 24	9 / 10	10 / 9	29 / 13	49 / 27	9 / 10	12 / 12
C / AM+Syn+C	18 / 176	30 / 188	15 / 177	21 / 187	18 / 176	37 / 197	15 / 177	26 / 198

- Category: categories of the item in question

  A: products, B: companies, C: persons, D: facilities, E: sights, F: information resources, G: online shops, H: events

- Specialty: the extent to which a hypothetical user knows the item in question

  A: detail, B: outline, C: difference from other items, D: little knowledge

Details of these subcategories are described in the overview paper by the organizers [10].

We analyze the evaluation results obtained by AM+Syn+C, which was most effective in Table 4.2, on a subcategory-by-subcategory basis. Tables 3 and 4 show the DCG and WRR of AM+Syn+C for TYPE=A and TYPE=AB, respectively. The column "#Topics" denotes the number of topics for each subcategory.

The column "Linked(%)" denotes the proportion of topics for which at least one relevant document was linked from another page in the Web collection. The values in this column is useful for analysis purposes, because our system highly depends on the anchor text that links to relevant documents. However, there was no significant difference between subcategories in terms of the values of the "Linked" column.

Because each topic can be classified into one or more subcategories for "Category", the total number of topics in "Category" is greater than the total number of topics used for the formal run. In Tables 3 and 4, "TYPE" and "Type" are different and should no be confused.

For "Type", the DCG and WRR for "Type 1" were greater than those for "Type 2" and "Type 3". Thus, in the navigational Web retrieval, it is crucial whether or not the information need can precisely be represented by a single keyword or phrase.

For "Category", the DCG and WRR for "B" and "H" were greater than those for the other subcategories. Thus, representative pages of products and companies can be retrieved with a high accuracy. The WRR for "C" was smaller than those for the other subcategories, while the DCG for "C" was comparable with those for most of the subcategories.

While the DCG is a cumulation of the scores for the relevant documents in the top ten documents, the WRR is calculated using only the first relevant documents found in the top ten documents. Thus, the WRR decreases rapidly as the rank of the first relevant documents decreases. In summary, it is still difficult to retrieve the representative page of a person with a high accuracy, when compared with other item subcategories.

For "Specialty", the DCG and WRR for "B" and "C" were greater than those for "A" and "D", although it is expected that a person who knows the target item in detail can represent an effective query. One reason is that the anchor-based retrieval, which contributes to the effectiveness of our system significantly, uses the anchor text produced by a large number of "general people". In other words, in topics "B" and "C" query terms are possibly similar to terms in the anchor text linking to relevant documents.

For example, the query of Topic 1063, which was categorized into "A" for the Specialty, is "yahoo housing information". However, the phrase "yahoo real estate" was used in most of the anchor texts linking to the relevant documents and "housing information" was not used.

To improve the retrieval accuracy for the "D" topics, we need to transform a user query into a more specific keyword. For example, the query of Topic 1167 is "Honda, bipedal robot", although the user produced this topic requires the information of "ASIMO". The retrieval accuracy was significantly improved when the term "ASIMO" was used as an alternative query. An automatic method for the query transformation needs to be explored.

5 Conclusion

In the Navi-2 subtask at the NTCIR-5 WEB Task, we used multiple methods to improve the retrieval accuracy. First, we improved the anchor text model. Second, we extracted synonyms from anchor text and expanded queries using those synonyms. Finally, we combined the anchor-based and content-based retrieval methods. Although the improvement obtained by each enhancement was small, when used together the improvement was noticeable.

Table 3. Evaluation results of AM+Syn+C for each topic subcategory (TYPE=A).

Subcategory		#Topics	Linked(%)	DCG-3-0	DCG-3-2	WRR-1-0	WRR-1-1
Type	1	145	96.6	3.101	3.565	0.767	0.797
	2	96	86.5	2.033	2.543	0.446	0.548
	3	28	85.7	1.383	1.657	0.356	0.388
Category	A	49	89.8	2.256	2.840	0.540	0.632
	B	60	95.0	3.071	3.386	0.717	0.740
	C	29	86.2	2.376	2.919	0.517	0.604
	D	29	79.3	2.502	3.113	0.637	0.706
	E	16	81.2	2.206	2.763	0.649	0.685
	F	47	97.9	2.403	2.806	0.555	0.586
	G	29	93.1	2.329	2.853	0.598	0.676
	H	19	100	3.117	3.607	0.768	0.851
Specialty	A	62	95.2	2.577	3.059	0.592	0.631
	B	106	92.5	2.720	3.262	0.632	0.711
	C	73	90.4	2.654	3.010	0.669	0.699
	D	28	85.7	1.594	1.986	0.435	0.508

Table 4. Evaluation results of AM+Syn+C for each topic subcategory (TYPE=AB).

Subcategory		#Topics	Linked(%)	DCG-3-0	DCG-3-2	WRR-1-0	WRR-1-1
Type	1	166	89.2	2.709	3.163	0.670	0.715
	2	112	82.1	1.739	2.293	0.381	0.507
	3	30	83.3	1.281	1.580	0.330	0.372
Category	A	59	79.7	1.874	2.404	0.448	0.533
	B	67	88.1	2.745	3.026	0.641	0.661
	C	33	78.8	2.078	2.634	0.452	0.547
	D	34	76.5	2.125	2.743	0.541	0.639
	E	17	76.5	2.054	2.579	0.605	0.641
	F	52	96.2	2.169	2.681	0.501	0.581
	G	34	88.2	1.980	2.577	0.508	0.642
	H	22	95.5	2.676	3.148	0.659	0.745
Specialty	A	76	82.9	2.102	2.634	0.483	0.565
	B	111	91.0	2.593	3.154	0.602	0.692
	C	89	85.4	2.175	2.535	0.548	0.596
	D	32	78.1	1.380	1.754	0.377	0.449

Acknowledgments

The authors would like to thank the organizers of the NTCIR-5 WEB task for their support with the Web collection.

References

[1] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks*, 30(1–7):107–117, 1998.

[2] N. Craswell, D. Hawking, and S. Robertson. Effective site finding using link anchor information. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 250–257, 2001.

[3] K. Eguchi, K. Oyama, E. Ishida, N. Kando, and K. Kuriyama. Overview of the Web retrieval task at the third NTCIR workshop. In *Proceedings of the Third NTCIR Workshop on Research in Information Retrieval, Automatic Text Summarization and Question Answering*, 2003.

[4] A. Fujii and T. Ishikawa. Japanese/English cross-language information retrieval: Exploration of query translation and transliteration. *Computers and the Humanities*, 35(4):389–420, 2001.

[5] A. Fujii and T. Ishikawa. Cross-language IR at University of Tsukuba: Automatic transliteration for Japanese, English, and Korean. In *Proceedings of the Fourth NTCIR Workshop on Research in Information Access Technologies Information Retrieval, Question Answering and Summarization*, 2004.

[6] M. Iwayama, A. Fujii, N. Kando, and Y. Marukawa. An empirical study on retrieval models for different document genres: Patents and newspaper articles. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 251–258, 2003.

[7] K. Järvelin and J. Kekäläinen. IR evaluation methods for retrieving highly relevant documents. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 41–48, 2000.

[8] J. Malawong and A. Rungsawang. Finding named pages via frequent anchor descriptions. In *Proceedings of the 11th Text REtrieval Conference*, 2002.

[9] K. Oyama, K. Eguchi, H. Ishikawa, and A. Aizawa. Overview of the NTCIR-4 WEB navigational retrieval task 1. In *Proceedings of the Fourth NTCIR Workshop on Research in Information Access Technologies Information Retrieval, Question Answering and Summarization*, 2004.

[10] K. Oyama, M. Takaku, H. Ishikawa, A. Aizawa, and H. Yamana. Overview of the NTCIR-5 WEB navigational retrieval subtask 2 (Navi-2). In *Proceedings of the Fifth NTCIR Workshop*, 2005.

[11] S. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 232–241, 1994.

[12] F. Smadja, K. R. McKeown, and V. Hatzivassiloglou. Translating collocations for bilingual lexicons: A statistical approach. *Computational Linguistics*, 22(1):1–38, 1996.

[13] K. Tanaka, A. Takasu, and J. Adachi. Finding named pages utilizing reliable title information. *IPSJ SIG Technical Report*, 2005-FI-78:17–24, 2005. (In Japanese).

[14] T. Westerveld, W. Kraaij, and D. Hiemstra. Retrieving Web pages using content, links, URLs and anchors. In *Proceedings of the 10th Text REtrieval Conference*, 2001.