Modeling Anchor Text and Classifying Queries to Enhance Web Document Retrieval

Atsushi Fujii
Graduate School of Library, Information and Media Studies
University of Tsukuba
1-2 Kasuga, Tsukuba, 305-8550, Japan
fujii@slis.tsukuba.ac.jp

ABSTRACT

Several types of queries are widely used on the World Wide Web and the expected retrieval method can vary depending on the query type. We propose a method for classifying queries into informational and navigational types. Because terms in navigational queries often appear in anchor text for links to other pages, we analyze the distribution of query terms in anchor texts on the Web for query classification purposes. While content-based retrieval is effective for informational queries, anchor-based retrieval is effective for navigational queries. Our retrieval system combines the results obtained with the content-based and anchor-based retrieval methods, in which the weight for each retrieval result is determined automatically depending on the result of the query classification. We also propose a method for improving anchor-based retrieval. Our retrieval method, which computes the probability that a document is retrieved in response to the given query, identifies synonyms of query terms in the anchor texts on the Web and uses these synonyms for smoothing purposes in the probability estimation. We use the NTCIR test collections and show the effectiveness of individual methods and the entire Web retrieval system experimentally.

Categories and Subject Descriptors

H.3.3 [**Information search and retrieval**]: Retrieval models

General Terms

Experimentation, Measurement

Keywords

Web retrieval, anchor text, query classification

1. INTRODUCTION

Recent research on document retrieval has shown that there are several types of queries on the World Wide Web and that the ideal retrieval methods differ fundamentally depending on the query type.

Broder [3] classified queries on the Web into the following three types.

WWW 2008, April 21–25, 2008, Beijing, China.
ACM 978-1-60558-085-2/08/04.

- navigational: the immediate intent is to reach a particular site
- informational: the intent is to acquire information assumed to be present on one or more Web pages
- transactional: the intent is to perform a Web-mediated activity

Through a questionnaire survey and user log analysis, Broder showed that there were many occurrences of each query type on the Web.

Existing test collections for Web retrieval, such as those produced for TREC and NTCIR, target the navigational and informational query types. Experimental results obtained with these test collections showed that the content of Web pages is useful for informational queries, whereas link or anchor information among Web pages is useful for navigational queries [5, 9, 14]. Li et al. [13] proposed a rule-based template-matching method to improve retrieval accuracy for transactional queries. Thus, classifying queries on the Web is crucial to selecting the appropriate retrieval method.

We propose methods to enhance Web retrieval and show their effectiveness experimentally. Our purpose is twofold. First, we propose a method to model anchor text for navigational queries. Compared with content-based retrieval, which has been studied for a long time, anchor-based retrieval has not been fully explored. Second, we propose a method to identify query types and use different retrieval methods depending on the query type. We target the navigational and informational query types because existing test collections do not target transactional queries.

Section 2 outlines our system for Web retrieval. Sections 3 and 4 elaborate on our anchor-based retrieval model and our method for classifying queries, respectively. Section 5 evaluates our methods and entire system experimentally.

2. SYSTEM OVERVIEW

Figure 1 shows the overall design of our Web retrieval system, which consists of three modules: "query classification", "content-based retrieval model", and "anchor-based retrieval model".

The purpose of our system is to produce a ranked document list in response to a query. We target informational and navigational queries.

For informational queries, the purpose is the same as in conventional ad-hoc retrieval. For navigational queries, a user knows a specific item (e.g., a product, company, or

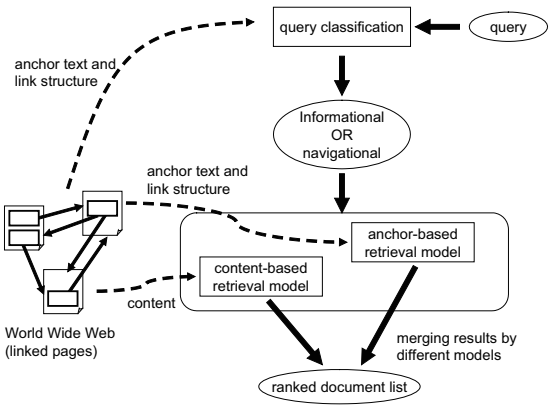

Figure 1: The overall design of our Web retrieval system.

person) and the purpose is to find one or more representative Web pages related to the item.

Irrespective of the query type, we always use both the content-based and anchor-based retrieval models. However, we change the weight of each model depending on the query type.

For preprocessing, we perform indexing for information on the Web. In content-based retrieval, index terms are extracted from the content of Web pages. In anchor-based retrieval, the anchor text and link structure on the Web are used for indexing purposes. We also use the anchor text and link structure to produce a query classifier.

We use the term "link" to refer to a hyperlink between two Web pages and the term "anchor text" to refer to a clickable string in a Web page used to move to another page. The following example is a fragment of a Web page that links to `http://www.acm.org/`. Here, "ACM Site" is anchor text.

```
<A HREF="http://www.acm.org/">ACM Site</A>
```

Given a query, we first perform query classification to categorize the query as informational or navigational. We then use both the content-based and anchor-based retrieval models to produce two ranked document lists, in each of which documents are sorted according to the score with respect to the query. Finally, we merge the two ranked document lists to produce a single list.

Because the scores computed by the two retrieval models can potentially have different interpretations and ranges, it is difficult to combine them in a mathematically sound way. Thus, we rerank each document by a weighted harmonic mean of the ranks in the two lists. We compute the final score for document d, $S(d)$, as follows.

$$S(d) = \frac{\alpha}{R_c(d)} + \frac{1-\alpha}{R_a(d)} \quad (0 \leq \alpha \leq 1) \tag{1}$$

$R_c(d)$ and $R_a(d)$ are the ranks of d in the content-based and anchor-based lists, respectively. α, which ranges from 0 to 1, is a parametric constant to control the effects of $R_c(d)$ and $R_a(d)$ in producing the final list. In brief, for informational queries, α should be greater than 0.5 so that $R_c(d)$ becomes more influential than R_a.

For the three modules in Figure 1, we use an existing model for content-based retrieval, but propose new methods for anchor-based retrieval (Section 3) and query classification (Section 4).

For the content-based retrieval, we index the documents in the Web collection by words and bi-words. We remove HTML tags from the documents and use ChaSen[1] to perform morphological analysis and extract nouns, verbs, adjectives, out-of-dictionary words, and symbols as index terms. We use Okapi BM25 [16] to compute the content-based score for each document with respect to a query.

We also perform pseudo-relevance feedback, for which we collect the top 10 documents in the initial retrieval and use the top 10 terms to expand the original query. The ranks of the terms are determined by the weight of each term.

3. ANCHOR-BASED RETRIEVAL MODEL

3.1 Overview

A number of methods have been proposed to use links and anchor text in Web retrieval.

Yang [19] combined content-based and link-based retrieval methods. To use link information for retrieval purposes, Yang used an extension of the HITS algorithm [11], which determines hubs and authoritative pages using a link structure on the Web. However, Yang did not use anchor text for retrieval purposes.

Craswell et al. [4] used anchor text as surrogate documents and used Okapi BM25, which is a content-based retrieval model, to index the surrogate documents, instead of the content of the target pages.

Westerveld et al. [18] also used anchor text as surrogate documents. However, because their retrieval method was based on a language model, they used the surrogate documents to estimate the probability of term t given surrogate document d, $P(t|d)$.

In Sections 3.2–3.4, we explain our anchor-based retrieval model. In Section 5.2, we compare the effectiveness of existing models and our model.

3.2 Entire Model

To use anchor text for retrieval purposes, we index the anchor text in a 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, $P(d|q)$. The task is to select the d that maximizes $P(d|q)$, which is transformed using Bayes' theorem as follows.

$$\arg\max_d P(d|q) = \arg\max_d P(q|d) \cdot P(d) \tag{2}$$

We have two alternative methods to estimate $P(d)$. First, we can use maximum likelihood estimation, which estimates $P(d)$ as the probability that d is linked via 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. Second, we can use PageRank [2], which estimates the probability that a user surfing the Web visits document d, $P(d)$. In Section 5.2, we compare the effectiveness of these two methods in estimating $P(d)$. To compute $P(q|d)$, we assume that the terms in q are independent and approximate

[1] http://chasen.naist.jp/hiki/ChaSen/

$P(q|d)$ as follows.

$$P(q|d) = \prod_{t \in q} P(t|d) \quad (3)$$

To extract term t in q, we use ChaSen to perform morphological analysis on q and extract nouns, verbs, adjectives, out-of-dictionary words, and symbols as index terms. We elaborate on two alternative models to compute $P(t|d)$ in Section 3.3.

We extract 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 discard the anchor text used to link to pages in the same server. Because we use a string matching method to identify servers, variants of the name of a single server, such as alias names, are considered different names. Additionally, even if a page links to another page more than once, we use only the first anchor text.

Because anchor texts are usually shorter than documents, 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 in resolving this problem. However, for navigational queries, the precision is usually more important than the recall. Thus, 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 follows.

$$\begin{aligned} P(t|d) &= P(t|s,d) \cdot P(s|d) \\ &\approx P(t|s) \cdot P(s|d) \end{aligned} \quad (4)$$

$P(t|s)$ denotes the probability that s is replaced with t. To derive the second line of Equation (4), 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 those of $P(t|d)$, which is explained in Section 3.3. We elaborate on the methods for extracting synonyms and computing $P(t|s)$ in Section 3.4.

However, if no synonyms of t are modeled in our system, a different smoothing method is necessary; otherwise the product calculated by Equation (3) becomes zero. For smoothing purposes, we replace $P(t|d)$ with $P(t)$, which is the probability that a term randomly selected from the anchor texts in 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 (3) is small. However, if mismatched query terms are low-frequency words, which are usually effective for retrieval purposes, $P(q|d)$ decreases substantially.

3.3 Modeling Anchor Text

To compute $P(t|d)$ in Equation (3), we use two alternative models.

In the first model, taken from Westerveld et al. [18], the set of all anchor texts linking to d, $\mathbf{A}_d$, is used as a single document, D, which is used as the 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. We call this the "document model".

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 follows.

$$P(t|d) = \sum_{a \in \mathbf{A}_d} P(t|a) \cdot P(a|d) \quad (5)$$

$P(t|a)$ 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 to the total frequency of all anchor texts in $\mathbf{A}_d$. To improve the efficiency of the computation for Equation (5), we consider only as that include t. We call this the "anchor model".

We illustrate the difference between these two models by 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 as informative as the 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, according to Equation (5), 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 between these two models with a hypothetical example. We use the top page of "Yahoo! Japan" (http://www.yahoo.co.jp/) as d and assume that d is linked from the following three anchor texts.

- $a_1 = \{\text{Yahoo, Japan}\}$
- $a_2 = \{\textit{yafuu}\}$
- $a_3 = \{\text{Yahoo}\}$

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.

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

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

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

Unlike the document model, $P(\textit{yafuu}|d)$ in the anchor model is greater than $P(\text{Japan}|d)$. In the real world, "*yafuu*" is more effective than "Japan" to search for "Yahoo! Japan".

The difference of these two models is also associated with spam. In the document model, the distribution of ts in $\mathbf{A}_d$ is

biased by a large anchor text linking to d and consequently the computation of $P(t|d)$ can be manipulated by an individual or a small group of people. In other words, the document model is vulnerable to spam.

However, the anchor model, which computes $P(t|d)$ on an anchor-by-anchor basis, is robust against spam. In Equation (5), $P(t|d)$ becomes large when t frequently appears in a and a frequently links to d. If many people use a to link to d, $P(a|d)$ becomes large, but it is difficult for an individual to manipulate $P(t|a)$. If an individual produces a that rarely links to d, he/she can manipulate $P(t|a)$, but $P(a|d)$ becomes small.

In summary, the anchor model is robust against spam and more intuitive than the document model. We compare the effectiveness of these two models quantitatively in Section 5.2.

3.4 Extracting Synonyms

When multiple anchor texts link to the same Web page, they 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 anchor texts to produce a link to "http://www.google.co.jp/".

While existing methods to extract translations use documents as a bilingual corpus [17], 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 may be 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 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 $P(t|s)$ that will be used in Equation (4)

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

In the second step, for any pair 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 a transliteration method [7]. If either e or j can be transliterated into its counterpart, 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 follows using Bayes' theorem.

$$\arg\max_t p(t|s) = \arg\max_d p(s|t) \cdot p(t) \tag{6}$$

$p(s|t)$ denotes the probability that t is transformed into s on a phoneme-by-phoneme basis. If $p(s|t) = 0$, t is not a transliteration of s. $p(t)$ denotes the probability that t is generated as a word in the target language [7]. However, in our case we always set $p(t) = 1$, because our purpose is to check whether or not two given words comprise a transliteration pair.

We extract (e,j) as a transliteration equivalent pair only if $p(e|j)$ or $p(j|e)$ takes a positive value. Because 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 (4), because we require the probability that t can substitute for s when used in an anchor text. Thus, Equation (6) is used only for extracting transliteration equivalents.

In the final step, we compute $P(t|s)$ as follows.

$$P(t|s) = \frac{F(t,s)}{\sum_{r \neq s} F(r,s)} \tag{7}$$

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

4. QUERY CLASSIFICATION

4.1 Overview

The purpose of query classification is to categorize queries into the informational and navigational types. A number of methods have been proposed [1, 9, 12].

Kang and Kim [9] used multiple features for query classification purposes and demonstrated their effectiveness using the TREC WT10g collection. Search topics for the topic relevance and homepage finding tasks were used as informational and navigational queries, respectively.

Lee et al. [12] performed human subject studies and showed that user-click behavior and anchor–link distribution are effective for query classification purposes. They also argued that the features proposed by Kang and Kim are not effective for query classification purposes. However, because they did not perform Web retrieval experiments, the effects of their query classification method on retrieval accuracy are not clear.

In this paper, we enhance the classification method proposed by Lee et al.and show its effectiveness in Web retrieval. However, we do not use user-click behaviors because we do not have search log information. We use only the anchor–link distribution, which can be collected from the anchor texts and link information in a Web collection. Thus, unlike a log-based query classification method [1], our method does not require large amounts of search log information.

In Section 5.3, we compare the effectiveness of our classification method and existing methods.

4.2 Methodology

The idea of the use of the anchor–link distribution proposed by Lee et al. [12] is as follows. For a navigational query, a small number of authoritative pages usually exist. Thus, the anchor text that is the same as the query is usually used to link to a small number of pages. However, for an informational query, the anchor text that is the same as the query, if it exists, is usually used to link to a large number of pages.

Given a query, Lee et al. computed its anchor–link distribution as follows. First, they located all the anchors appearing on the Web that had the same text as the query, and extracted their destination URLs. Then, they counted how many times each destination URL appeared in this list and sorted the destinations in the descending order of their appearance. They created a histogram in which the frequency

count in the ith bin was the number of times that the ith destination appeared. Finally, they normalized the frequency in each bin so that the frequency values summed to one. Figure 2 shows example histograms produced by this method, in which (a) and (b) usually correspond to histograms for navigational and informational queries, respectively. While the distribution in (a) is skewed, the distribution in (b) is uniform.

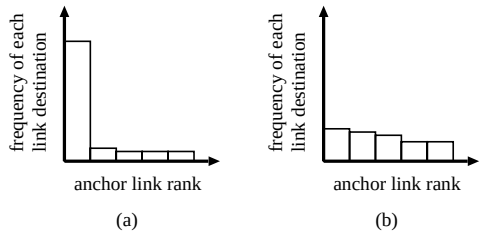

Figure 2: Example histograms for anchor–link distribution.

To distinguish the two histograms depicted by Figures 2 (a) and (b), Lee et al. computed how skewed the distribution was, for which several standard statistical measures were used.

However, Lee et al. considered only anchor texts that were exactly the same as the given query. Thus, if a given query consisted of more than one term, such as "information retrieval" and "trec, nist, test collection", and there was no anchor text exactly the same as this query, the anchor–link distribution for this query could not be computed. This limitation is also problematic for queries consisting of a single term, if a query is in an agglutinative language, in which multiple query terms are combined without lexical segmentation.

To resolve this problem, we modify Lee et al.'s method. In brief, if a query does not appear in the anchor texts in the Web collection as it is, we decompose the query into terms and compute the anchor–link distribution for each term.

We consider a set of terms in query q, $\mathbf{T}_q$. We also consider a set of documents linked by the anchor texts including $t \in \mathbf{T}_q$, $\mathbf{D}_t$. We use ChaSen to extract terms t from query q. To quantify the degree to which the anchor–link distribution for q is skewed, unlike Lee et al.'s method, we compute the conditional entropy of $\mathbf{D}_t$ given $\mathbf{T}_q$, $H(\mathbf{D}_t|\mathbf{T}_q)$ as follows.

$$H(\mathbf{D}_t|\mathbf{T}_q) = -\sum_{t\in\mathbf{T}_q} P(t)\cdot\sum_{d\in\mathbf{D}_t} P(d|t)\cdot\log P(d|t) \qquad (8)$$

If the anchor–link distribution for each t is skewed, $H(\mathbf{D}_t|\mathbf{T}_q)$ becomes small. If the anchor–link distribution for each t is close to uniform, $H(\mathbf{D}_t|\mathbf{T}_q)$ becomes large. If all terms in a query are used together in the same anchor text, $H(\mathbf{D}_t|\mathbf{T}_q)$ tends to become small.

$P(t)$ denotes the probability with which term t appears in query q. Because queries are usually short, we use the uniform distribution of t and thus $P(t) = \frac{1}{|\mathbf{T}_q|}$.

$P(d|t)$ denotes the probability that document d is linked by the anchor texts including term t. $P(d|t)$ is the length of the bin including d in the histogram produced by Lee's method. In Figure 2, each bin denotes the frequency of destination documents linked by a specific anchor text, divided by the total frequency of all documents in the histogram. While Lee et al. assumed that q and t were identical and considered only the distribution of $P(d|t)$, we assume that q consists of more than one term and consider a combination of $P(d|t)$ for different ts.

Using $H(\mathbf{D}_t|\mathbf{T}_q)$, we compute the degree to which query q should be regarded as an informational query, $i(q)$. We divide $H(\mathbf{D}_t|\mathbf{T}_q)$ by $\log|\mathbf{D}_t|$, so that the range of the value of $i(q)$ is $[0,1]$.

$$i(q) = \frac{H(\mathbf{D}_t|\mathbf{T}_q)}{\log|\mathbf{D}_t|} \qquad (9)$$

If $i(q)$ is less than 0.5, we determine that q is a navigational query; otherwise we determine that q is an informational query.

We have two alternative methods for using $i(q)$. First, we use $i(q)$ only to determine the query type. The value of α in Equation (1) is determined independently. Second, we use $i(q)$ as α in Equation (1), so that we can determine the value of α automatically. In Section 5, we compare the effectiveness of these methods.

We can further enhance our classification method. If term t is not included in the anchor texts on the Web, we use a synonym of t to compute $i(q)$. To extract a synonym of a term, we use the method proposed in Section 3.4. However, we simply replace t with s and do not use $P(t|s)$ in the computation of $i(q)$.

In summary, we have resolved three issues that were not addressed in Lee et al. [12]. First, our method can compute the anchor–link distribution of queries for which the query text does not exist as anchor text on the Web. Second, our method can determine the weight of the content-based and anchor-based retrieval methods automatically. Finally, our method can use synonyms of query terms for smoothing purposes.

Our method is associated with two parametric constants. $i(q)$ can potentially be small, if few of terms in q are used in the anchor texts on the Web. In such a case, q is regarded as a navigational query irrespective of the informational need behind q. To avoid this problem, if term t is not used in the anchor texts, we estimate the frequency of documents linked by anchor text including t by a default value. We empirically set this parameter to 10 000. The other parameter is the bin size in the histogram as in Figure 2. We empirically set this parameter to 5. The values of these parameters should be determined by the size of a target Web collection. We have not identified an automatic method to determine the optimal values. However, this issue is also related to Lee's method.

5. EVALUATION

5.1 Evaluation Method

We evaluated the effectiveness of our proposed methods with three experiments. First, we evaluated the effectiveness of the anchor-based retrieval model proposed in Section 3. Second, we evaluated the effectiveness of the query classification method proposed in Section 4. Finally, we evaluated the accuracy of our Web retrieval system as a whole, proposed in Section 2.

Table 1 shows a summary of the test collections used for our experiments. We use the test collections produced for NTCIR-3 [6] and NTCIR-4 [5, 14]. These share a target doc-

ument set, which consists of 11 038 720 pages collected from the JP domain. Thus, most of the pages are in Japanese. The file size is approximately 100 GB, which is 10 times the size of the TREC WT10g collection.

Table 1: Test collections used for experiments.

	NTCIR-3	NTCIR-4		NTCIR-5
Topic type (#Topics)	info (47)	info (80)	navi (168)	navi (841)
Doc size	100 GB			1 TB
Avg#Rels	75.7	84.5	1.79	1.94
Avg#Terms	2.89	2.39	1.39	1.35
Experiments	Sections 5.3, 5.4, 5.5			
			Section 5.2	

We also used the test collection produced for NTCIR-5 [15]. The target document set for NTCIR-5 consists of 95 870 352 pages collected from the JP domain. The file size is approximately 1 TB, which is 10 times the size of the NTCIR-3/4 collection.

Search topics are also in Japanese. While the NTCIR-3 collection includes only informational search topics, the NTCIR-4 collection includes both informational and navigational search topics. Because these topics target the same document collection, we can use them to evaluate our query classification. However, because the NTCIR-5 collection includes only navigational search topics, we use these topics to evaluate the anchor-based retrieval model.

In the relevance judgment, the relevance of each document with respect to a topic was judged as "highly relevant", "relevant", "partially relevant", or "irrelevant". We used only topics for which at least one highly relevant or relevant document was found. As a result, we collected 47 topics from the NTCIR-3 collection and 80 informational topics and 168 navigational topics from the NTCIR-4 collection, respectively, and a further 841 topics from the NTCIR-5 collection. Thus, we used 1009 (168 + 841) topics for the evaluation of the anchor-based retrieval model, and 127 (47 + 80) informational topics and 168 navigational topics for the evaluation of the query classification.

We used the highly relevant and relevant documents as the correct answers. The average numbers of correct answers were 75.7 for the NTCIR-3 information topics, 84.5 for the NTCIR-4 informational topics, 1.79 for the NTCIR-4 navigational topics, and 1.94 for the NTCIR-5 navigational topics, respectively.

For each topic, we used only the terms in the "TITLE" field, which consists of one or more terms, as a query. The average numbers of terms were 2.89 for the NTCIR-3 information topics, 2.39 for the NTCIR-4 informational topics, 1.39 for the NTCIR-4 navigational topics, and 1.35 for the NTCIR-5 navigational topics, respectively.

In Section 5.2, we evaluate the anchor-based retrieval methods, for which we used only the navigational queries in the NTCIR-4 and NTCIR-5 collections. In this evaluation, we used Mean Reciprocal Rank (MRR) as the evaluation measure. MRR has commonly been used to evaluate precision-oriented retrieval, such as retrievals for navigational queries and question answering. For each query, we calculated the reciprocal of the rank at which the first correct answer was found in the top 10 documents. MRR is the mean of the reciprocal ranks for all queries.

In Section 5.3, we evaluate the accuracy of query classification methods, for which we used both the informational and navigational queries in the NTCIR-3 and NTCIR-4 collections. We also evaluate the contribution of each query classification method to the retrieval accuracy. We used Mean Average Precision (MAP) and MRR as the evaluation measures of the retrieval accuracy. MAP, which considers both precision and recall, is appropriate to evaluate the retrieval for the informational queries. To calculate MAP, we used the top 100 documents.

In Section 5.4, we analyze the errors of our query classification method and their effects on the retrieval accuracy, for which we used the NTCIR-3 and NTCIR-4 collections.

In Section 5.5, we evaluate our system as a whole. We used the NTCIR-3 and NTCIR-4 collections and evaluated a combination of the methods proposed in this paper.

5.2 Evaluating the Anchor-based Retrieval Model

Using the 168 navigational queries in NTCIR-4 and the 841 navigational queries in NTCIR-5, we compared the MRR of the following retrieval methods.

- CC: a content-based retrieval model that uses Okapi BM25 to index the content of the target pages (Section 2)
- CS: a content-based retrieval model that uses anchor texts as surrogate documents and uses Okapi BM25 to index them [4]
- ADP: an anchor-based retrieval model (Section 3) that computes $P(t|d)$ and $P(d)$ by the document model [18] and PageRank, respectively
- ADM: the same as ADP but computes $P(d)$ by the maximum likelihood estimation
- AAP: an anchor-based retrieval model that computes $P(t|d)$ and $P(d)$ by the anchor model proposed in Section 3.3 and PageRank, respectively
- AAM: the same as AAP but computes $P(d)$ by the maximum likelihood estimation
- AAMS: a combination of AAM and the synonym-based smoothing proposed in Section 3.4
- AAMSC: a combination of CC and AAMS according to Equation (1)

Table 2 shows the MRR for these retrieval methods. The relative superiority between the two methods was almost the same for the NTCIR-4 and NTCIR-5 test collections.

Comparison of CC and CS, which used the same retrieval model but indexed different information, shows that the use of the anchor text was effective in substantially improving MRR.

Comparison of CS and each of the anchor-based model variations ADP, ADM, AAP, AAM, AAMS, and AAMSC, which used the same information but used different models, shows that the method of modeling anchor text was crucial. For navigational queries, our anchor-based retrieval model was more effective than Okapi BM25, irrespective of the implementation variation.

Comparison of ADP and ADM (or AAP and AAM), which used the same retrieval model but different implementations

Table 2: MRR of retrieval methods for navigational queries.

Method	NTCIR-4	NTCIR-5
CC	0.063	0.047
CS	0.458	0.446
ADP	0.556	0.556
ADM	0.590	0.675
AAP	0.567	0.577
AAM	0.606	0.691
AAMS	0.612	0.691
AAMSC	0.618	0.691

for $P(d)$, shows that maximum likelihood estimation was more effective than PageRank in the computation of $P(d)$.

Comparison of ADP and AAP (or ADM and AAM), which used the same retrieval model but used different implementations for $P(t|d)$, shows that the anchor model proposed in this paper was more effective than an existing method [18].

Comparison of AAM and AAMS shows that synonym-based smoothing was effective in improving MRR in NTCIR-4. Through a topic-by-topic analysis, we found that the improvement was caused by topic #0064, for which an English translation of the query is "The Princeton Review of Japan"[2]. For this query, the reciprocal rank was 0 without smoothing. However, the reciprocal rank was 1 with smoothing.

Comparison of AAMS and AAMSC shows that combining the anchor-based and content-based retrieval models was effective in improving MRR in NTCIR-4, but not in NTCIR-5. The optimal value of α was determined by preliminary experiments. We set $\alpha = 0.3$ and $\alpha = 0.1$ for NTCIR-4 and NTCIR-5, respectively.

In summary, a) the anchor text model, b) the smoothing method using automatically extracted synonyms, and c) a combination of the anchor-based and content-based retrieval models were independently effective in improving the accuracy of navigational Web retrieval.

Because the above items a) and b) were proposed in this paper, our contribution improved MRR from 0.590 (ADM) to 0.606 (AAM) and 0.612 (AAMS) for NTCIR-4, and from 0.675 (ADM) to 0.691 (AAM/AAMS) for NTCIR-5. We used the paired t-test for statistical testing, which investigates whether the difference in performance is meaningful or simply because of chance [8, 10]. The differences of ADM and AAM for NTCIR-4 and NTCIR-5 were significant at the 5% and 1% levels, respectively. However, the differences between AAM and AAMS were not significant for NTCIR-4 and NTCIR-5. We can conclude that the anchor text model was effective in improving the accuracy of navigational Web retrieval.

5.3 Evaluating Query Classification

Using the 127 informational queries and the 168 navigational queries in NTCIR-3 and NTCIR-4, we evaluated the effectiveness of our query classification method.

As comparisons, we used the query classification methods proposed by Kang and Kim [9] and Lee et al. [12]. Kang's method used four features: distribution of query terms, mutual information, usage rate as an anchor text, and part-of-speech information. Kang and Kim integrated the four features by a linear combination, for which the optimal weights of each feature were determined manually. However, because manual optimization of different weights was prohibitive, we evaluated each feature independently.

We did not use the part-of-speech feature. Because the TREC queries used by Kang and Kim included natural language phrases, verbs appear in informational queries more often than in navigational queries. However, because the NTCIR queries consist of only nouns, it is obvious that the part-of-speech feature is not effective for query classification purposes.

For each of the remaining three features, we implemented a query classifier. Each classifier computes the score of a given query and determines the query type by comparing the score and a predetermined threshold. Because the threshold of each classifier must be determined manually, we evaluated each classifier using different values of the threshold and selected the optimal value.

Kang and Kim used two threshold values and did not identify the query type if the score fell between the two threshold values. This method is effective in improving accuracy, although it decreases coverage. However, because manual optimization of different threshold values was prohibitive, we used a single threshold for each classifier.

First, we compared the following methods in terms of the accuracy of query classification.

- DI: the distribution of query terms feature in Kang and Kim's method
- MI: the mutual information feature in Kang and Kim's method
- AN: the usage rate of anchor text in Kang and Kim's method
- LM: Lee's method
- OM: our method

While for MI and AN, we set the threshold to 0, for DI we set the threshold to 0.6. For LM and OM, we set the threshold to 0.5.

Table 3 shows the accuracy of different query classification methods. It is apparent that in Kang and Kim's method, the accuracy of AN was greatest. The accuracy of OM was greater than those of the other methods. Thus, our query classification method was more effective than these existing methods.

Table 3: Accuracy of query classification methods.

Method	Accuracy
DI	53.9
MI	43.1
AN	75.6
LM	72.5
OM	79.3

We analyzed the effect of synonym-based expansion and found that the following two queries, which are both navigational queries in the NTCIR-4 collection, were correctly classified by the synonym-based expansion: #0010 "SHARP, liquid crystal TV" and #0078 "France, sightseeing".

[2]The Princeton Review of Japan is an educational institution. http://www.princetonreview.co.jp/index_e.html

Second, we compared the following methods in terms of the accuracy of Web retrieval.

- NC: no query classification, with α in Equation (1) always 0.5 irrespective of the query type
- AN: the usage rate of anchor text in Kang and Kim's method
- LM: Lee's method
- O1: our method, with α predefined
- O2: our method, with α determined automatically by Equation (9)
- CT: correct query type defined in the NTCIR-3/4 collections

Because NC is a baseline method, any method with accuracy smaller than that of NC has no utility. For AN, LM, O1, and CT, α was 0.7 for informational queries and 0.3 for navigational queries, respectively. These values of α were determined by preliminary experiments. However, because O2 used the value of $i(q)$ as α, manual optimization was not required.

Each method used CC and AAMS in Section 5.2 for the content-based and anchor-based retrieval models, respectively. Thus, the MAP and MRR of each method were determined only by the query classification accuracy.

Because Kang and Kim did not compare their method with the case of no classification (NC), our experiment is the first effort to evaluate the contribution of query classification to Web retrieval accuracy.

Table 4 shows the MAP and MRR for the different retrieval methods. Although CT outperformed the other methods in MAP and MRR, our methods (O1 and O2) outperformed NC, AN, and LM in MAP and MRR. Thus, our query classification method was more effective in improving Web retrieval accuracy than the existing automatic classification methods.

In the existing classification methods, AN outperformed NC and LM. We used the paired t-test for statistical testing and found that the difference between AN and each of our methods (O1 and O2) was significant at the 5% level in MAP but was not significant in MRR. However, in Section 5.5 we show that a combination of our proposed methods improved the MAP and MRR of a baseline retrieval system significantly.

Table 4: MAP and MRR of retrieval methods for informational and navigational queries.

Method	MAP	MRR
NC	0.254	0.468
AN	0.281	0.504
LM	0.265	0.485
O1	0.300	0.519
O2	0.304	0.517
CT	0.312	0.545

5.4 Error Analysis for Query Classification

We analyzed the queries that were misclassified by our method (OM in Table 3). We also analyzed how the retrieval accuracy was changed by the errors, for which we compared O2 and CT in Table 4 with respect to AP (Average Precision) and RR (Reciprocal Rank). Note that MAP and MRR are evaluation measures for all queries and that for each query only AP and RR can be calculated.

We identified two error types for informational queries and four error types for navigational queries. Table 5 shows the number of cases and changes of AP and RR for each error type. In Table 5, "↓", "=", and "↑" denote "decrease", "equality", and "increase" of AP/RR for O2 compared with those for CT. Although AP and RR were usually decreased by misclassified queries, for some queries AP or RR were increased by the classification error.

Table 5: Error types of query classification and changes of AP and RR.

Error type	Query type	#Errors	AP			RR		
			↓	=	↑	↓	=	↑
(a)	info	14	14	0	0	10	3	1
(b)	info	9	9	0	0	5	3	1
(c)	navi	27	8	9	10	6	16	5
(d)	navi	1	1	0	0	1	0	0
(e)	navi	4	0	4	0	0	4	0
(f)	navi	1	1	0	0	1	0	0

In the following, we elaborate on the error types (a)–(f). To exemplify queries, we use English translations of original Japanese queries. While errors (a)–(c) occurred because we decomposed a query into more than one term, errors (d)–(f) were common to query classification methods that use anchor texts on the Web.

Error (a). As explained in Section 4.2, if there is no anchor text that is the same as a query, we decompose the query into more than one term. When these terms are used as independent anchor texts, the entropy for each term is small and $i(q)$ for this query is also small. An example query is #0001 "offside, football, rule" in the NTCIR-4 collection.

There were 14 queries misclassified for this reason. For 10 queries RR decreased and for three queries RR did not change. However, for one query, #0112 "sauna, Finland" in the NTCIR-4 collection, RR increased.

Error (b). When all terms in a query appear in the same anchor text, $i(q)$ for this query becomes small. An example query is #0029 "photoshop, tips" in the NTCIR-4 collection.

There were nine queries misclassified for this reason. For five queries RR decreased and for three queries RR did not change. However, for one query, #0013 "*Kyoto*, temple, shrine" in the NTCIR-3 collection, RR increased from 0.33 to 1. In this example, although a user intended to submit an informational query, a representative page related to sightseeing for *Kyoto* was found by the anchor-based retrieval module.

Error (c). Like error (a), this error occurs because we decompose the query into more than one term. However, unlike error (a), because these terms were general words that

were frequently used in different anchor texts, the entropy for each term was large and $i(q)$ was also large.

An example query is #0159 "Venusline, sightseeing" in the NTCIR-4 collection. In this example, because "Venusline" is a proper noun of a road, this term tends to be used as a navigational query. However, the entropy of "sightseeing" was so large that the entropy of "Venusline" was overshadowed. As a result, $i(q)$ for this query became large.

Error (d). When no terms in a query appear in the anchor texts in the Web collection, $i(q)$ for the query becomes large. This case often happens when a query consists of an infrequent proper noun, such as #0160 "*Ganryuu* island" in the NTCIR-4 collection.

Error (e). Query #0092 "genetically modified food" appears in a number of anchor texts in different contexts, such as "The homepage of genetically modified food at the Ministry of Health, Labour, and Welfare" and "Frequently asked questions for genetically modified food". Therefore, the distribution of the destination documents linked by these anchor texts was not skewed. As a result, $i(q)$ for this query became large.

Error (f). If the degree of skewness is not sufficient, $i(q)$ becomes greater than 0.5. There was only a single such query, #0192 "Coca-Cola", for which $i(q)$ was 0.549.

5.5 Evaluating the Entire Retrieval System

We evaluated the accuracy of our Web retrieval system as a whole. As shown in Figure 1, our system consists of three modules. For each module, a baseline system used the existing method that achieved the highest accuracy in our experiments: ADM in Table 2 for the anchor-based retrieval model and AN in Table 3 for the query classification. All systems used CC in Table 2 as the content-based retrieval model.

Table 6 shows the MAP and MRR of the different retrieval systems, in which BL denotes the results of the baseline system. The results for O1, O2, and CT are the same as those in Table 4. In Table 6, our systems (O1 and O2) outperformed the baseline system in both MAP and MRR.

Table 6: MAP and MRR of retrieval systems.

System	MAP	MRR
BL	0.272	0.491
O1	0.300	0.519
O2	0.304	0.517
CT	0.312	0.545

Table 7 shows the results of the paired t-test for statistical testing, in which "$<$" and "$\ll$" indicate that the difference of two results was significant at the 5% and 1% levels, respectively, and "—" indicates that the difference of two results was not significant. The difference between BL and O1 was statistically significant for MAP and MRR. The difference between BL and O2 was also statistically significant for MAP and MRR. The difference between CT and each of our systems (O1 and O2) was not significant in MAP.

In summary, irrespective of whether the value of α is determined manually or automatically, our system outperformed the baseline system significantly in MAP and MRR. Thus, we can reduce the manual cost required to optimize the value of α. In addition, our proposed methods significantly improved the accuracy of Web document retrieval.

Table 7: t-test results for differences between retrieval systems ("$\ll$": 0.01, "$<$": 0.05, "—": not significantly different).

	MAP	MRR
BL vs. O1	$\ll$	$<$
BL vs. O2	$\ll$	$<$
O1 vs. CT	—	$\ll$
O2 vs. CT	—	$\ll$

6. CONCLUSION

There are several types of queries on the Web and the expected retrieval method can vary depending on the query type. We have proposed a Web retrieval system that consists of query classification, anchor-based retrieval, and content-based retrieval modules.

We have proposed a method to classify queries into the informational and navigational types. Because terms in navigational queries often appear in the anchor text of links to other pages, we analyzed the distribution of query terms in the anchor texts on the Web for query classification purposes. While content-based retrieval is effective for informational queries, anchor-based retrieval is effective for navigational queries. Our retrieval system combines the results obtained with the content-based and anchor-based retrieval methods, in which the weight for each retrieval result is determined automatically depending on the result of the query classification.

We have also proposed a method to model anchor text for anchor-based retrieval. Our retrieval method, which computes the probability that a document is retrieved in response to a given query, identifies synonyms of query terms in the anchor texts on the Web and uses these synonyms for smoothing purposes in the probability estimation.

We used the 100 GB and 1 TB Web collections produced in NTCIR workshops, and showed the effectiveness of individual methods and the entire Web retrieval system experimentally. Our anchor-based retrieval method improved the accuracy of existing methods. In addition, our entire system improved the accuracy of the baseline system. These improvements were statistically significant.

Although we targeted the informational and navigational queries, future work includes targeting other types of queries, such as transactional queries.

7. ACKNOWLEDGMENTS

The author would like to thank the organizers of the NTCIR WEB task for their support with the Web test collections. This research was supported in part by MEXT Grant-in-Aid Scientific Research on Priority Area of "New IT Infrastructure for the Information-explosion Era" (Grant No. 19024007).

8. REFERENCES

[1] R. Baeza-Yates, L. Calderón-Benavides, and C. González-Caro. The intention behind Web queries. In *Proceedings of the 13th International Conference on String Processing and Information Retrieval*, pages 98–109, 2006.

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

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

[4] 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.

[5] K. Eguchi, K. Oyama, A. Aizawa, and H. Ishikawa. Overview of the WEB task at the fourth NTCIR workshop. In *Proceedings of the Fourth NTCIR Workshop on Research in Information Access Technologies Information Retrieval, Question Answering and Summarization*, 2004.

[6] 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.

[7] 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.

[8] D. Hull. Using statistical testing in the evaluation of retrieval experiments. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 329–338, 1993.

[9] I.-H. Kang and G. Kim. Query type classification for Web document retrieval. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 64–71, 2003.

[10] E. M. Keen. Presenting results of experimental retrieval comparisons. *Information Processing & Management*, 28(4):491–502, 1992.

[11] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. In *Proceedings of the ninth annual ACM-SIAM symposium on Discrete algorithms*, pages 668–677, 1998.

[12] U. Lee, Z. Liu, and J. Cho. Automatic identification of user goals in Web search. In *Proceedings of the 14th International World Wide Web Conference*, pages 391–400, 2005.

[13] Y. Li, R. Krishnamurthy, S. Vaithyanathan, and H. V. Jagadish. Getting work done on the Web: Supporting transactional queries. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 557–564, 2006.

[14] 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.

[15] 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 Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access*, pages 423–442, 2005.

[16] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. In *Proceedings of the 3rd Text REtrieval Conference*, 1994.

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

[18] 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.

[19] K. Yang. Combining text- and link-based retrieval methods for Web IR. In *Proceedings of the 10th Text REtrieval Conference*, pages 609–618, 2001.