Research Paper

The test collection for navigational retrieval on WWW data—Design and characteristics

Keizo Oyama[1], Haruko Ishikawa[2], Koji Eguchi[3], Akiko Aizawa[4]

[1, 2, 3, 4] *National Institute of Informatics (NII)*

[1, 3, 4] *The Graduate University for Advanced Studies (SOKENDAI)*

ABSTRACT

This paper describes the design and characteristics of a test collection for navigational retrieval of WWW data that was built through the WEB Task of the Fourth NTCIR Workshop to evaluate the retrieval effectiveness of Web search systems. This reusable test collection consists of 100 gigabytes of Web document data and 300 topics of various types and corresponding relevance judgments. Among the several types of 'Navigational Retrieval,' we selected the 'Known Item Search,' which simulates a situation where a user searches for one or a few 'representative Web pages' of a known item. It is assumed that the user knows about the item but may not have seen its Web page. Relevance judgments were performed on the probable documents mainly from the viewpoint of representativeness of respective known items represented by the topics. Using the judgment results, several evaluation measures were applied to various retrieval results. Based on the evaluation results, relationships among the types of topics, Web-page styles and search methods are discussed. The stability of the evaluation results with different numbers of topics is also analyzed.

KEYWORDS

Web information retrieval, evaluation methods, test collections

1 Introduction

This paper describes the design, construction process and characteristics of a test collection for navigational retrieval on the WWW data to evaluate the retrieval effectiveness of Web search systems.

The test collection was built through the Navigational Retrieval Task 1 that was conducted from 2002 to 2004 as a subtask of the WEB Task at the Fourth NTCIR Workshop (NTCIR-4 WEB) [1].

Several kinds of tasks can be associated with the term "Navigational Retrieval". We selected "Known Item Search" as the first task to tackle. Thus, we call this subtask "Navigational Retrieval Task 1." In this task, we attempted to evaluate the retrieval effectiveness of Web search systems from the viewpoint of "Known Item Search," in which a user searches for one or a few "representative Web pages" for an item about which the user already knows (this concept is described in more detail in Section 2.2).

The test collection consists of 100 gigabytes of Web document data (NW100G-01), 300 topics, and the corresponding relevance judgments. It was built in the Navigational Retrieval Task 1 using the following process: (1) the document data was constructed (we used the same document data as the NTCIR-3 WEB [3,4]); (2) 300 topics were created by 11 people; (3) the document data and the topics were distributed to nine participants, and, in turn, 16 search run results were submitted by five groups and some others by the organizers; (4) relevance judgment on the run results were performed by 10 assessors; and (5) 218 topics, each of

Received November 22, 2004; Revised February 2, 2005; Accepted February 21, 2005.

{1)oyama, 2)haruko, 3)eguchi, 4)akiko}@nii.ac.jp

which had one or more relevant documents in the document data, were selected for system evaluation. The resulting test collection is thus reusable.

Relevance judgments were performed, by one assessor for each topic, on both the documents pooled from the results and those found by following probable hyperlinks or by searching possible URL patterns.

Each run result submitted by the participants was evaluated using relevance judgments with several different measures.

Similar tasks have been conducted in TREC. One was the "Home/Named Page Finding Task" [9] in the TREC-2003 Web Track. Its purpose was to evaluate the effectiveness of systems to search for a mixture of home pages and named pages by their names.

The "Known Item Search" is different in that one or a few search terms (not necessarily a name) are provided to specify a searched item, rather than a name of a Web page. Therefore, there may be several independent relevant pages. Moreover, a relevant page may be a single page or a top page of a closely interlinked page group. It is considered to reflect the real search scene more appropriately.

In the following, we describe the task definition in Section 2, the document set in Section 3, the search topics in Section 4, run conditions in Section 5, relevance assessment in Section 6, system evaluation results in Section 7, topic and relevant document analysis in Section 8, stability analysis in Section 9 and we give our conclusions in Section 10.

2 Task Definition

2.1 Search target items

An item that can be a search target is a "known item" that represents a specific thing or a matter, or a collection of specific things or matters. Searches on unspecific things or matters or on unspecific information for information gathering purposes are not handled in this test collection.

Some search target items do not actually exist on the Web, such as products, organizations, stores, persons, facilities, natural things and events, whereas other search target items do exist on the Web, such as information-providing sites, search services, data files and documents. Although general information cannot be a search target, information that has a specific content and is assumed to be provided in a "representative Web page" can be a search target.

2.2 Known items

An item is regarded as "known" when a searcher knows beforehand by some means that the search target item exists and can identify the item if search result pages are presented.

However, as in the following examples, the searcher may not be able to describe the item exactly to specify it.

- Knows only an acronym
- Cannot express the item clearly with a few words or phrases
- Remembers its features but has forgotten its name

On the other hand, the item's "representative Web page" itself may not necessarily be "known" and may fall into any of the following three cases:

- Has seen the page and remembers its outline.
- Has actually seen the page but does not clearly remember what the page was like.
- Has never seen the page but anticipates that such a page exists.

2.3 Representative Web pages

We propose that a "representative Web page" for "a known item" should satisfy the conditions described below. However, they are presented to clarify the nature of the task, and are strictly neither necessary nor sufficient conditions.

(1) Provider of the Web page
The provider must be an organization or a person that is responsible for the "known item" or an organization or a person that is generally appreciated as an authority on the "known item".

(2) Content of the Web page
The Web page must cover information that is strongly and comprehensively related in all aspects with the "known item" as far as it is provided by the Web page provider. The page should also include as little information not directly related to the item as possible. The "strongly related information" may either be given in the Web page or in an easily recognized link from the Web page.

3 Document Data Set

The document data set that is searched is NW100G-01, which was also used in the "NTCIR-3 WEB Test Collection [2]." It consists of text files of approximately 100 gigabytes in total and their metadata, which were crawled from http servers in the "*.jp" domain from August to November in 2001.

Search results can include two types of documents as follows:

(1) Crawled pages: Web pages included in NW100G-01; and

(2) Referred pages: Web pages that have at least one link from Web pages in (1) but are not included in (1), and are actually fetched and stored in referral storage for relevance judgment purposes.

Document identifiers of all these Web pages are given in a file provided with the test collection. A file including the list of links from documents in (1) to documents in (1) and (2) is also provided. Refer to the references [3] and [4] for more details.

We use documents only in group (1) for the analysis in this paper.

4 Search topics

4.1 Topic creation policy

It is important that the topics used in the test collection reflect real information needs in information retrieval on the Web.

One method to realize this is to sample queries from existing search engine logs. In the 'Large Web task' of Trec-8 and 9 [5], [6], topics consisted of a minimum of two non-stop words in a query for which assessors 'felt' they understood the original intention of the searchers and were able to judge the relevance of documents. Demerits for this method are that: (1) it is generally difficult to obtain search engine logs; (2) there will be arbitrariness in the selection of the topics; (3) such information as searchers' attributes are usually unavailable and it is difficult to interpret a searcher's intention; and (4) the queries may be influenced by the characteristics of the search engine used.

Another unique method was used for the 'home page finding task' of TREC 10 [7], the 'Named page finding' task of TREC 2002 [8] and the Navigational task (which combines the home page and named page finding tasks) of TREC 2003 [9]. Relevant (or 'answer') Web pages were first collected from the .GOV document collection of TREC, which were extracted only from Web sites in the "gov" domain. Then, for the home page finding task, the topics were derived from the name of the site while for the Named page finding task, topics were derived from the name of the page. The two types of topics were arbitrarily mixed for the navigational task. Demerits for this method are that: (1) the limitation of the site coverage may not agree with the actual search needs; (2) the method might be appropriate to sample the searchable relevant pages but is not equivalent to sampling real information needs; and (3) creating topics to find known Web pages is not suitable for a situation where unknown Web pages (though the search objects are known) are to be searched.

The most commonly used method in creating test collections is to have actual searchers create topics reflecting real information needs. The problem of biased topics is usually dealt by having as many topic creators as possible and by asking them to think of realistic search scenarios. Furthermore, because the properties of each topic creator are known, if they are provided with the topics, future analysis can be done accordingly. This approach has the merit that, by having the topic creators judge the relevance of documents, the accuracy of assessment can be high. We took this approach.

4.2 Creation and selection procedure

The test collection includes 300 topics for evaluating search effectiveness. We selected them from 456 topics that were first created by 11 topic creators, discarding similar and inappropriate ones from several viewpoints. Most of the topic creators were undergraduate or graduate students of several universities.

To make the topics reflect real information needs, they used the following procedure:

(1) each topic creator recollected a natural search target item in relation to hobbies, study, work, or daily life,

(2) the creator imagined a corresponding "representative Web page", and

(3) wrote the search target item and a phrase that expresses the Web page in a free format.

(4) Organizers selected those appropriate as "known items" from (3).

(5) Each topic creator described it as a search topic in a given format.

In the creation and selection process, it was not checked whether relevant documents existed in the document data set. However, because the document data set was collected between August and November, 2001, items for which representative Web pages could not have existed at that time were excluded from the search topics.

4.3 Elements of search topics

Each search topic consists of the elements described below, which were designed to be used for testing search methods and analyzing topic characteristics, both depending on searchers' attributes and types of information need. The original language is Japanese but English translations are also available, mainly for publication purposes[1].

(1) NUM: Topic number

A topic number used as topic id.

(2) TYPE: Topic type

A topic type code defined as follows:

1: A single search term specifies the known item;

2: A combination of search terms specifies the known item;

3: A single search term or a combination of

[1] Sample topics of various types and categories are shown in Appendix A.

search terms represents the known item but cannot specify it.

(3) CATEGORY: Category of the known item

One or more category codes for the known item, defined as below.

A: Products / services

B: Companies / organizations (including shops and administrative organs)

C: Persons

D: Facilities (including public and private)

E: Sights and historic spots, and natural things (including parks)

F: Information resources (including information sites and data files)

G: Events

Z: Others

(4) TITLE: Search Terms

Search terms supposed to be submitted to a search engine to meet the user's information needs: up to three terms (or phrases) in the order of importance. Called TITLE for historical reasons.

(5) DESC: Description

A sentence briefly describing the information need; should be conceptually consistent with the search terms.

(6) NARR: Narrative of the information need

Explanation of the information need. All the following subelements are optional.

(6-1) NARR/TERM: Explanation of terms

Sentences describing definition of meanings and explaining related terms regarding terms in Search Terms and Description when they have ambiguity or they are not popular.

(6-2) NARR/BACK: Explanation of background

Sentences explaining background of the information need and the motivation.

(6-3) NARR/RELE: Relevance criteria

Sentences explaining relevance criteria on the item and the pages when they are not clear just with TITLE and DESC.

(7) USER: Searcher's Attributes

Position, sex, and experience years of Web search.

(7-1) USER/SPECIALTY: Searcher's knowledge level

A code denoting searcher's knowledge level on the searched item defined as follows:

A: Knows the item in detail.

B: Knows the outline of the item.

C: Knows the item to the extent the item can be identified from others.

D: Knows existence of the item but very little about it.

5 Run conditions

In the NTCIR-4 WEB Task, participants were requested to execute search runs using the following combinations of topic elements but not using other topic elements[2].

(1) TITLE only (mandatory)

(2) Any combination of TITLE, DESC, and NARR/BACK

(3) Any combination of TYPE and CATEGORY added to (1)

(4) Any combination of TYPE and CATEGORY added to (2)

Each participant submitted run results and a system description form. Each run result included up to 100 retrieved documents for each topic. The system description form included a concise description of each run including the items below among others that were used to analyze topic characteristics.

Topic Part

The topic elements used for the search run

Query Method

Automatic or interactive

Query Expansion

Techniques used to expand queries

Link Information

Use of link information

Anchor

Use of anchor text

Users of the test collection may use any elements for experiments. It is desirable however to make evaluations based on comparison with (1) as a baseline.

6 Relevance assessment

Relevance assessment was performed on documents included in 16 run results submitted by five participants and 68 run results added by the authors to find as many relevant documents as possible. Each run result included up to 100 documents for each topic.

Pooling was applied to the run results for the relevance assessment. However, we requested assessors to find relevant documents as far as possible by following possible hyperlinks and by searching probable URLs. That means any documents included in the overall document data set, not just the documents in the pools or the documents in the run results, potentially become the targets of relevance assessment. While we expect that most of the documents have been assessed, it is yet to be verified.

Although we tried to have relevance assessment of each topic done by the topic creator, only about half of the topics were actually so treated.

[2] No run using combination (3) or (4) was actually submitted.

On completion of the assessment, all of the relevance judgment results were inspected by the organizers, and a few topics that did not meet the judgment criteria set by the organizers were carefully re-judged.

Thus, we are confident that the test collection can be considered reusable.

6.1 Judgment basis

The assessors were requested to use as the judgment basis not only text but also clues that the assessors usually used in Web browsing and which searchers in general are assumed to use, such as page titles, host names, URL patterns and various kinds of HTML tags.

For frame set pages and pages that automatically jump to other URLs, the assessor refers to their link target pages as far as they are included in the document data set and takes them into the judgment basis.

Further procedural instructions were not given to the assessors, as we consider that the relevance of the search results should be judged with a subjective view of each assessor as a general searcher.

6.2 Relevance judgment

Relevance of each document to the search topic was assigned to one of the following levels by absolute evaluation:

A: Relevant

A representative page appropriate for the searched item satisfying the information need.

B: Partially relevant

A page that partially satisfies the information need, as follows:

- ◆ A representative page of an item covering a superordinate or subordinate concept of the searched item, or a page covering only part of the searched item; an explicit hyperlink to the relevant document should be provided in the page;
- ◆ A page that can be regarded as a substitute for the representative page of the searched item.

D: Non-relevant

Otherwise.

Judgment of the representativeness of the Web pages, including the provider and the contents, relies on the expertise and the common sense of each assessor.

6.3 Duplicate pages

When relevant or partially relevant pages had identical entities or were corresponding pages within mirror sites, judging from their contents and URL's and so on, these pages were judged as duplicate pages. Even when the content text was completely the same, pages that were considered to have different link target pages or images were not deemed to be duplicate pages.

When a pair of relevant or partially relevant pages were linked with an explicit anchor, the link source page may be deemed as a duplicate page of the link destination page. In this case, when the link source page appears in the run result, the link destination page is also deemed to have appeared. However, the reverse does not hold.

The judgment results of duplicate pages are to be utilized to investigate how duplicate pages affect evaluation of search effectiveness in the future.

7 System evaluation methods

The test collection is designed so that systems can be evaluated on four combinations of the document data sets and the relevance levels described below.

<u>Document data sets</u>

(DS-1) Document data set consisting of crawled pages and referred pages

The document data set defined in Section 3. It consists of not only documents with page data in the NW100G-01, but also documents that have only in-links from one or more of the stored documents and were actually fetched.

(DS-2) Document data set consisting of crawled pages only

An additional document data set for comparison. It consists of only the documents with page data contained in NW100G-01.

<u>Relevance levels</u>

(RL-1) Rigid:

Documents assessed as 'relevant' are regarded as relevant documents.

(RL-2) Relaxed:

Documents assessed as 'relevant' and 'partially relevant' are regarded as relevant documents.

<u>Search topics</u>

Although 300 topics are provided in the test collection, only topics that satisfy the following conditions are meaningful for system evaluation:

Condition 1: at least one relevant document at the 'rigid' relevance level was found or was considered to exist in the Web space when the documents were crawled; and

Condition 2: at least one relevant document at the 'relaxed' relevance level was found in the document set.

Consequently, the following two topic sets were used for system evaluation with the test collection:

(TS-1): 218 topics for (DS-1); and
(TS-2): 197 topics for (DS-2).

In the NTCIR-4 WEB Task, each run was evaluated on all four combinations of data sets and relevance levels. However, many run results submitted by the participants did not include referred pages because their systems could retrieve only documents with content text.

For the current paper, only the combination of document set DS-2, relevance level RL-1 and topic set TS-2 is used, to evaluate the selected runs on a common basis because of the above-mentioned limitation.

In the task, DCG (Discounted Cumulative Gain) and MRR (Mean Reciprocal Rank)[3] [3], [4] at the top-ranked 100 document level are used as the evaluation measures.

Although many topics have multiple relevant documents, most of them are redundant, i.e., either duplicate or closely linked Web pages. Therefore, for such a group of pages, the top ranked relevant document has some importance and the others have little.

Because duplication and link relations are not considered in the evaluation, the appropriateness of DCG values as the system effectiveness should be investigated further. However, because only the first relevant document retrieved is used in MRR, the appropriateness is the same regardless of considering duplication or link relations. Consequently, we use MRR here.

8 Summary of topics and relevant document characteristics

As described in Section 4.2, topics in the test collection are characterized in three types (TYPE 1 to 3) and eight categories (CATEGORY A to Z). In this section, we look at the detail of 168 topics for which the relevant documents were found by 10 selected runs through the task[4].

The number of topics and their proportions as percentages (in brackets) for each type and category in the 168 topics are listed in Table 1.

In Figure 1, the numbers of topics and their types versus the numbers of systems that were unable to find the relevant documents for the topics are shown. The leftmost bars of Figure 1 show that there were 13 topics in total for which all 10 runs were able to find the relevant documents. Among those 13 topics, 11 were Type 1, and the remaining two topics were Types 2 and 3 respectively. The rightmost bars of the figure show the number and the types of topics for which all 10 runs were unable to find relevant documents. The figure also shows a tendency that fewer runs were able to find relevant documents for topic Type 2.

Table 1. Proportion of ‘TYPE’ and ‘CATEGORY’ in topics

CATEGORY	ALL	TYPE 1	TYPE 2	TYPE 3
ALL	168(100%)	116(69.0%)	46(27.4%)	6(3.57%)
A	24(14.3%)	8(33.3%)	16(66.7%)	0(0%)
B	45(26.8%)	39(86.7%)	5(11.1%)	1(2.2%)
C	15(8.9%)	10(66.7%)	4(26.7%)	1(6.6%)
D	21(12.5%)	16(76.2%)	4(19.0%)	1(4.8%)
E	17(10.1%)	14(82.3%)	2(11.8%)	1(5.9%)
F	18(10.7%)	8(44.4%)	7(38.9%)	3(16.7%)
G	7(4.2%)	4(57.1%)	3(42.9%)	0(0%)
Z	2(1.2%)	1(50.0%)	1(50.0%)	0(0%)
Combo	19(11.3%)	15(78.9%)	4(21.1%)	0(0%)

Note: The percentages in the TYPE columns denote their proportions in respective CATEGORY rows.

Compositions of categories in the same topics in Figure 1 are shown in Figure 2.

More systems had difficulties in finding relevant documents for Category A (Products/services) than B (Companies/organizations). This may be because of the fact that the search terms specified in TITLE of Category B topics are more likely to be specific about the representative Web pages while the search terms of Category A topics are also likely to appear on pages unrelated to the topic. Furthermore, it is interesting to note that Category B topics are likely to specify site top pages while relevant documents of Category A topics could appear lower in the hierarchy of the Web site.

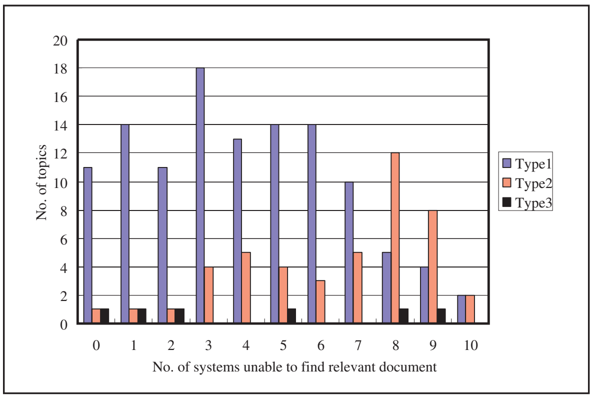


Fig. 1 Types of retrieved topics

[3] Definitions of DCG and MRR are given in Appendix C.

[4] The 10 selected runs are described in Section B-2 of Appendix B.

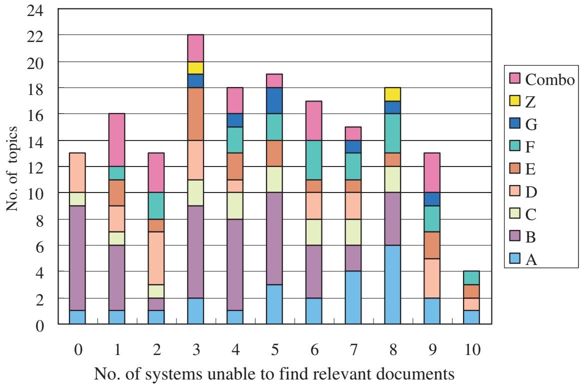


Fig. 2 Categories of retrieved topics

Fewer systems found relevant documents also for Category F (Information resources). This may also be because the search terms of Category F topics are less likely to specify the representative Web pages directly; in addition, there are more Type 2 topics in Category F in comparison to other categories.

We further investigated the distribution of types and categories of topics in the following groups, listed in Table 2.

(a) Topics for which all systems returned relevant documents within the top 10.

(b) Topics for which all systems returned relevant documents within the top 100.

(c) Topics for which relevant documents were returned within the top 10 by all systems except one.

(d) Topics for which only systems that used anchor text information could return relevant documents within the top 10.

(e) Topics for which none of the systems could return relevant documents.

Types and categories of search topics and their search terms, the URL and title of the most frequently linked relevant page and the number of in-links to the page are shown in Table 2. It should be noted that defining the topic difficulty may be a complex task, especially in navigational retrieval, as the retrieval results may be directly influenced by the actual style of the relevant Web pages, regardless of the topics.

Topics of groups (a) and (b) are mostly Type 1 and Category B. The most frequently linked relevant documents of the 13 topics were linked on average 100 times from documents within the delivered data set. However, relevant documents of group (a) are linked less than 10 times, yet these two topics have the higher rank when averaging over all runs. Relevant documents for these two topics have relatively simple structure and the exact search terms appear in TITLE elements and META tags. Relevant documents of topics in group (c) where all systems performed well except for one system that performed poorly for the topic are similar to those for the two topics in group (a). Relevant documents are moderately linked within the delivered data set and the exact search terms appear in TITLE elements and META tags. There were no FRAMESET pages in groups (a) to (c).

For the topics in group (d), relevant documents were retrieved easily only by runs that utilized anchor text information. The relevant documents that belong to this group are linked frequently and many were FRAMESET pages or pages displayed by FLASH.

For the first topic in group (e), not only the 10 selected runs but all runs (described in Appendix B) except two link-based runs provided by organizers, failed to return the relevant documents. These two runs did not rely on anchor text information but used the link information from the content base search results. For the latter topic of group (e), more link-based runs provided by organizers were able to find relevant documents though they performed poorly.

9 Stability of evaluation results in terms of number of topics

One of the important attributes of a good test collection for various information retrieval systems is that it contains a sufficient number of topics for the evaluation. For topical search, Voorhees and Buckley [14] evaluated the effect of topic set size on retrieval experiments by estimating the ‘error rates’ when determining a system’s superiority with the mean average precision. The maximum topic set size they calculated was 25, because of the difference in the nature of the search types and the number of relevant documents available. In the Navigational Retrieval task, we consider mean reciprocal rank (MRR)[5] to be the most appropriate evaluation measure. To demonstrate that we provided a sufficient number of topics, we calculated the consistency of system ranking determined by the mean reciprocal rank over topic sets of different sizes, using the system evaluation results described in Appendix B. Of the 168 topics, we randomly selected 25 (forming five topic groups), 50 (forming three topic groups), 75 (forming two topic groups) and 100 (forming two topic groups, topics partly overlapping). No group in a given topic set size contains the same topics, except for the topic set size 100. Then we calculated Spearman and Kendall correlation coefficients respectively for the system rank determined from each topic size group and that determined from 168 topics.

[5] Definition of MRR is described in Appendix C.

Table 2 Topic and Web page characteristics of relevant documents

	Type	Search terms	URL of most frequently linked relevant page	Page Title	In-links
(a)	1, B	関東自動車学校 (Kanto Motor School)	http://www.ktds.co.jp/	関東自動車学校ホームページ	9
	1, C	鷲見一夫 (Kazuo Sumi)	http://researchers.adm.niigata-u.ac.jp/030003010120_a.html	新潟大学研究者総覧：法学部鷲見一夫	2
(b)	1, A	はとバス (Hato Bus)	http://www.hatobus.co.jp/	HATOBUS ONLINE	243
	1, B	可視化情報学会 (The Visualization Society of Japan)	http://www.vsj.or.jp/indexe.html	Visualization Society	180
	1, B	日本オリンピック委員会 (Japanese Olympic Committee)	http://www.joc.or.jp/index.html	--日本オリンピック委員会---Japanese Olympic Committee---	257
	1, B	ヤマトヤシキ (Yamatoyashiki)	http://www.yamatoyashiki.co.jp/	ヤマトヤシキのホームページ	11
	1, B	アエロフロート (Aeroflot)	http://www.aeroflot.gr.jp/jp/index2.htm	AEROFLOT Japan	62
	1, B	銀行業務検定協会 (The Association of Banking Business Proficiency Test)	http://www.kenteishiken.gr.jp/index.htm	銀行業務検定協会	74
	3, B	福助 , 株式会社 (Fukusuke, corporation)	http://www.fukusuke.co.jp/	福助株式会社	90
	1, B	加古川商工会議所 (East Harima Katsumesian's Union)	http://www.kakogawa-cci.or.jp/	そ加古川商工会議所ホームページへ	17
	1, D	日本科学未来館 (National Museum of Emerging Science and Innovation)	http://www.miraikan.jst.go.jp/	National Museum of Emerging Science and Innovation	113
	2, D	飛鳥資料館 , 奈良文化財研究所 (Asuka Historical Museum, Nara Cultural Properties Research Institute)	http://www.asukanet.gr.jp/asukahome/index.html	ASUKA/Home Page	99
	1, D	新国立劇場 (New National Theatre)	http://www.nntt.jac.go.j	新国立劇場ホームページ	133
(c)	1, D	地下鉄博物館 (Tokyo Subway Museum)	http://www.tokyometro.go.jp/ttjin/5200.html	地下鉄博物館	50
	1, E	登呂遺跡 (The Toro remains)	http://www.city.shizuoka.shizuoka.jp/torohaku/sub2.htm	登呂遺跡について	7
	1, F,A	JA長野県農業情報ネットワーク (JA Nagano Agricultural Information Network)	http://www.janis.or.jp/agrinet/	JA長野県農業情報ネットワーク	31
(d)	2, A	新幹線，700 系 (Shinkansen, 700 series)	http://www.jr-central.co.jp/museum/zukan/sin_700.html (FLASH)	JR東海museum	35
	1, A	コカコーラ (Coca-Cola)	http://www.cocacola.co.jp/ (FLASH)	Welcome to Coca-Cola Japan	166
	1, B	巨人 (Giant)	http://giants.yomiuri.co.jp/ (FRAME)	東京読売巨人軍	1894
	1, B	伊勢丹百貨店 (ISETAN Department Store)	http://www.isetan.co.jp/	ISETAN	1078
	1, B	フィンランド (Finland)	http://www.finland.or.jp/index-j.html (FRAME)	フィンランド大使館，東京	163
	1, D	国立演芸場 (National Engei Hall)	http://www.ntj.jac.go.jp/	国立劇場	40
	2, F	ゴルフ，ルール (Golf, rules)	http://www.golf-gtpa.or.jp/rule_book/index.html (FRAME)	* There is no title element.	18
	1, F	日本経済新聞 (Nihon Keizai Shimbun)	http://www.nikkei.co.jp/	Nikkei Net	8772
(e)	1, E	吉野ケ里遺跡 (The Yoshinogari remains)	http://www.sagatokimeki.ne.jp/ (FRAME)	佐賀ときめき大学	16
	1, D	横浜国際競技場 (International Stadium Yokohama)	http://www.city.yokohama.jp/me/w-cup/stadium/index.html	横浜市　スポーツ　ワールドカップ	22

According to the average values that are shown for three topic set groups in Table 3, topic set sizes above 25 give relatively consistent results for the comparison of 10 runs.

Table 3 Correlation coefficients

No. topics	Kendall	Spearman
25	0.75	0.87
50	0.84	0.92
75	0.90	0.95
100	0.93	0.97

All correlations were significant, with p-values less than 1%, except for the 25-topic group, where the p-value is less than 5%.

10 Conclusion

In this paper, we described a test collection to evaluate navigational retrieval techniques on the Web, which was built through the Navigational Retrieval Task 1, a subtask of the WEB Task at the Fourth NTCIR Workshop. It was aimed at evaluating Web search engine systems for retrieving representative Web pages of known items.

The test collection consists of a 100-gigabyte Web document data set, NW100G-01, constructed at the Third NTCIR Workshop, 300 topics created at the Fourth NTCIR Workshop, and corresponding relevance judgments. Relevance assessment was done so that relevant documents were collected as comprehensively as possible. Consequently a reusable test collection was built. However, because the number of systems that submitted run results was not large enough and the variety of search methods was not sufficient, and also because the systems' detailed information is not available, users of the test collection can make only rough comparisons of the evaluation results of their system with those of the participants. For a strict analysis, they must execute comparative experiments by implementing comparable search methods themselves.

Relations between difficulty of topics and several attributes of topics and relevant documents were discussed. For instance, it was suggested that the topics having search terms that are likely to specify the representative Web pages and having a relevant document of the actual Web pages with many in-links and without frame structure are the easiest.

No specific details can be concluded on the sufficiency of the number of topics and further analysis is required. However, it seems that 168 topics gave a stable evaluation results.

Acknowledgments

This work was partially supported by the Grants-in-Aid for Scientific Research on Priority Areas of "Informatics" (#13224087) from the Ministry of Education, Culture, Sports, Science and Technology, Japan. We greatly appreciate the efforts of all the participants of the Navigational Retrieval Task 1 of the WEB Task at the Fourth NTCIR Workshop. We also appreciate the helpful advice of Professors Jun Adachi and Noriko Kando, National Institute of Informatics.

References

[1] Keizo Oyama, Koji Eguchi, Haruko Ishikawa, and Akiko Aizawa, "Overview of the NTCIR-4 WEB Navigational Retrieval Task 1," *Working Notes of the 4th NTCIR Workshop Meeting*, National Institute of Informatics, Tokyo, Japan, Supplement vol. 1, pp. ov-23-40, Oct. 2004. Available: http://research.nii.ac.jp/ntcir-ws4/NTCIR4-WN/WEB/NTCIR4WN-OV-WEB-B-OyamaK.pdf

[2] "Research Purpose Use of NTCIR Test Collections," National Institute of Informatics. Available: http://research.nii.ac.jp/ntcir/permission/perm-en.html

[3] Koji Eguchi, Keizo Oyama, Emi Ishida, Noriko Kando, and Kazuko Kuriyama, "Overview of the Web Retrieval Task at the Third NTCIR Workshop," *NII Tech. Rep.*, no. NII-2003-002E, Jan. 2003. Available: http://research.nii.ac.jp/TechReports/03-002E.html

[4] K. Eguchi, K. Oyama, E. Ishida, N. Kando, and K. Kuriyama, "Evaluation Methods for Web Retrieval Tasks Considering Hyperlink Structure," *IEICE Trans. Inf. & Syst.*, vol. E86-D, no. 9, pp. 1804-1813, 2003. Available: http://search.ieice.org/2003/files/e000d09.htm#e86-d,9,1804

[5] D. Hawking, Ellen M. Voorhees, N. Craswell, and Peter Bailey, "Overview of the TREC-8 Web Track," *8th Text Retrieval Conference (TREC-8)*, Gaithersburg, Maryland, Nov. 1999. Available: http://trec.nist.gov/pubs/trec8/t8_proceedings.html

[6] D. Hawking, "Overview of the TREC-9 Web Track," *Proc. TREC-9*, Gaithersburg, Maryland, Nov. 2000. Available: http://trec.nist.gov/pubs/trec9/t9_proceedings.html

[7] D. Hawking and N. Craswell, "Overview of the TREC-2001 Web Track," *Proc. TREC-2001*, Gaithersburg, Maryland, Nov. 2001. Available: http://trec.nist.gov/pubs/trec10/t10_proceedings.html

[8] N. Craswell and D. Hawking, "Overview of the TREC-2002 Web Track," *Proc. TREC-2002*, Gaithersburg, Maryland, Nov. 2002 Available: http://trec.nist.gov/pubs/trec11/t11_proceedings.html

[9] N. Craswell and D. Hawking: "Overview of the TREC 2003 Web Track,"*Proc. TREC-2003*, Gaithersburg, M aryland, Nov. 18-21, 2003. Available: http://trec.ni st.gov/pubs/trec12/papers/WEB.OVERVIEW.pdf

[10] Keizo Oyama, Koji Eguchi, Haruko Ishikawa, and Akiko Aizawa, "Overview of the NTCIR-4 WEB Navigational Retrieval Task 1," *Proc. 4th NTCIR Workshop*, 2004.

[11] Tomoyosi Akiba, Atsushi Fujii, Katunobu Itou, and Tetsuya Ishikawa, "Experiments on Web Retrieval Driven by Spontaneously Spoken Queries," *Proc. 4th NTCIR Workshop*, 2004.

[12] Teruhito Kanazawa, Tomonari Masada, Atsuhiro Takasu, and Jun Adachi, "R2D2 at NTCIR-4 Web Retrieval Task," *Proc. 4th NTCIR Workshop*, 2004.

[13] Hitoshi Nakakubo, Peng Zhang, and Takashi Sato, "NTCIR-4 WEB Experiments at Osaka Kyoiku University - Static/Dynamic Scoring Using Link Structure Analysis and Web Page Grouping," *Proc. 4th NTCIR Workshop*, 2004.

[14] Ellen M. Voorhees and Chris Buckley, "The Effect of Topic Set Size on Retrieval Experiment Error," *Proc. 25th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pp. 316-323, 2002.

[15] Namazu Project, "Namazu: a Full-text search Engine." Available: http://www.namazu.org/index.html.en

APPENDIX A. Samples of topics

List A (a) Sample topics of various TYPES

Original (in Japanese)	**Translation (in English)**
<NUM>0011</NUM> <TYPE>1</TYPE><CATEGORY>B</CATEGORY> <TITLE>DVD フォーラム</TITLE> <DESC>DVDフォーラムについて知りたい</DESC> <NARR> <TERM>DVDフォーラムとは、DVD の諸規格を定める組織である。</TERM> <BACK>近頃、DVD 規格の乱れが気になった。</BACK> </NARR> <USER SPECIALITY="A">大学院修士 1 年 , 男性 , 検索歴 4 年</USER>	<NUM>0011</NUM> <TYPE>1</TYPE><CATEGORY>B</CATEGORY> <TITLE>DVD Forum</TITLE> <DESC>Find information on DVD Forum.</DESC> <NARR> <TERM>DVD Forum is an organization that defines various DVD format specifications.</TERM> <BACK>Confusion about DVD specifications has been on my mind recently.</BACK> </NARR> <USER SPECIALITY="A">1st year Master's student, male, 4 years of searching experience.</USER>
<NUM>0017</NUM> <TYPE>2</TYPE><CATEGORY>F, A</CATEGORY> <TITLE>フォトショップ, tips</TITLE> <DESC>「フォトショップ」のちょっとしたテクニックを知りたい</DESC> <NARR> <TERM>「フォトショップ」はAdobe Photoshopとする。</TERM> <BACK>「フォトショップ」をもう少しうまく使いこなしたいので、「フォトショップ」のいろいろな小技が知りたい。</BACK> </NARR> <USER SPECIALITY="B">大学 3 年 , 女性 , 検索歴 3年</USER>	<NUM>0017</NUM> <TYPE>2</TYPE><CATEGORY>F, A</CATEGORY> <TITLE>Photoshop, tips</TITLE> <DESC>Search for documents with good tips for using "Photoshop."</DESC> <NARR> <TERM> "Photoshop" refers to Adobe Photoshop.</TERM> <BACK>I would like to learn some good tips for using "Photoshop" as I would like to master it a bit better.</BACK> </NARR> <USER SPECIALITY="B">3rd year undergraduate student, female, 3 years of searching experience.</USER>
<NUM>0093</NUM> <TYPE>3</TYPE><CATEGORY>F</CATEGORY> <TITLE>人口 , 日本 , 2000 年</TITLE> <DESC>2000 年の日本における人口を知りたい</DESC> <NARR> <BACK>社会調査の課題であった。</BACK> <RELE>公式統計情報のページを適合とする</RELE> </NARR> <USER SPECIALITY="A">大学院修士 1 年 , 男性 , 検索歴 3 年</USER>	<NUM>0093</NUM> <TYPE>3</TYPE><CATEGORY>F</CATEGORY> <TITLE>Population, Japan, year 2000</TITLE> <DESC>Find out the Japanese population in 2000.</DESC> <NARR> <BACK>This was an assignment on social surveys.</BACK> <RELE>Pages on official statistical information are relevant.</RELE> </NARR> <USER SPECIALITY="A">1st year Master's student, male, 3 years of searching experience.</USER>

List A (b) Sample topics of various CATEGORIES

Original (in Japanese)	**Translation (in English)**
<NUM>0010</NUM> <TYPE>2</TYPE><CATEGORY>A</CATEGORY> <TITLE>SHARP, 液晶テレビ</TITLE> <DESC>SHARPの液晶テレビの製品ラインナップを見てみたい。</DESC> <NARR></NARR> <USER SPECIALITY="C">大学2年, 男性, 検索歴3年</USER>	<NUM>0010</NUM> <TYPE>2</TYPE><CATEGORY>A</CATEGORY> <TITLE>SHARP, LCD TV</TITLE> <DESC>Find a product lineup of SHARP LCD TVs.</DESC> <NARR> </NARR> <USER SPECIALITY=“C”>2nd year undergraduate student, male, 3 years of searching experience.</USER>
<NUM>0012</NUM> <TYPE>1</TYPE><CATEGORY>B</CATEGORY> <TITLE>JPNIC</TITLE> <DESC>JPNICについて知りたい</DESC> <NARR> <TERM>JPNIC とは、日本のドメイン・IP などを管理する組織である。</TERM> <BACK>ドメイン管理の経緯に興味があった。</BACK> </NARR> <USER SPECIALITY="A">大学院修士1年, 男性, 検索歴4年</USER>	<NUM>0012</NUM> <TYPE>1</TYPE><CATEGORY>B</CATEGORY> <TITLE>JPNIC</TITLE> <DESC>Find information on JPNIC.</DESC> <NARR> <TERM>JPNIC is an organization that manages domains and IPs in Japan.</TERM> <BACK>I am interested in the history of/sequence of events leading up to domain management.</BACK> </NARR> <USER SPECIALITY="A">1st year Master's student, male, 4 years of searching experience.</USER>
<NUM>0086</NUM> <TYPE>3</TYPE><CATEGORY>C</CATEGORY> <TITLE>中村桂子</TITLE> <DESC>中村桂子が開設しているページを探したい。</DESC> <NARR> <RELE>生物学者である中村桂子が開設しているページが適合する</RELE> </NARR> <USER SPECIALITY="A">大学4年, 男性, 検索歴4年</USER>	<NUM>0086</NUM> <TYPE>3</TYPE><CATEGORY>C</CATEGORY> <TITLE>Keiko Nakamura</TITLE> <DESC>Search for pages set up by Keiko Nakamura.</DESC> <NARR> <RELE>Pages established by Keiko Nakamura, a biologist, are relevant.</RELE> </NARR> <USER SPECIALITY="A">4th year undergraduate student, male, 4 years of searching experience.</USER>
<NUM>0094</NUM> <TYPE>1</TYPE><CATEGORY>D</CATEGORY> <TITLE>国立オリンピック記念青少年総合センター</TITLE> <DESC>国立オリンピック記念青少年総合センターの情報を見たい</DESC> <NARR> <TERM>国立オリンピック記念青少年総合センターは代々木にある研修室や宿泊室を兼ね備えた施設である</TERM> <RELE> </RELE> </NARR> <USER SPECIALITY="A">大学院修士1年, 男性, 検索歴4年</USER>	<NUM>0094</NUM> <TYPE>1</TYPE><CATEGORY>D</CATEGORY> <TITLE>National Olympic Memorial Youth Center</TITLE> <DESC>Find information on the National Olympic Memorial Youth Center.</DESC> <NARR> <TERM>The National Olympic Memorial Youth Center is a facility with training rooms and accommodation located in Yoyogi.</TERM> <RELE> </RELE> </NARR> <USER SPECIALITY="A">1st year Master's student, male, 4 years of searching experience.</USER>
<NUM>0269</NUM> <TYPE>3</TYPE><CATEGORY>E</CATEGORY> <TITLE>安土城, 博物館</TITLE> <DESC>安土城跡を管理している博物館の情報が欲しい</DESC> <NARR></NARR> <USER SPECIALITY="B">大学2年, 男性, 検索歴5年</USER>	<NUM>0269</NUM> <TYPE>3</TYPE><CATEGORY>E</CATEGORY> <TITLE>Azuchi Castle, museum</TITLE> <DESC>Find information of the museum that maintains the Azuchi Castle ruins.</DESC> <NARR> </NARR> <USER SPECIALITY=“B”>2nd year undergraduate student, male, 5 years of searching experience.</USER>
<NUM>0096</NUM> <TYPE>2</TYPE><CATEGORY>F</CATEGORY> <TITLE>小泉純一郎, 所信表明演説, 官邸</TITLE> <DESC>小泉総理大臣の所信表明演説を見たい。</DESC> <NARR> <RELE>官邸内の大臣発言録のページを適合とする</RELE> </NARR> <USER SPECIALITY="A">大学2年, 男性, 検索歴5年</USER>	<NUM>0096</NUM> <TYPE>2</TYPE><CATEGORY>F</CATEGORY> <TITLE>Junichiro Koizumi, policy speech, official residence</TITLE> <DESC>Find articles on policy speeches made by Prime Minister Junichiro Koizumi.</DESC> <NARR> <RELE>Pages recording statements made by ministers in his official residence are relevant.</RELE> </NARR> <USER SPECIALITY="A">2nd year undergraduate student, male, 5 years of searching experience.</USER>
<NUM>0032</NUM> <TYPE>2</TYPE><CATEGORY>G</CATEGORY> <TITLE>土浦, 花火大会</TITLE> <DESC>土浦の花火大会について調べたい</DESC> <NARR> <BACK>全国の花火師が集まる大会だといわれているらしいが、どんな花火がどのくらい上げられているのか調べてみたいと思った</BACK> </NARR> <USER SPECIALITY="D">大学1年, 女性, 検索歴4年</USER>	<NUM>0032</NUM> <TYPE>2</TYPE><CATEGORY>G</CATEGORY> <TITLE>Tsuchiura, firework festival</TITLE> <DESC>Find information on Tsuchiura Fireworks Competition.</DESC> <NARR> <BACK>It is said that pyrotechnists gather for this festival from all over the country. I would like to know what kind and how many fireworks are set off.</BACK> </NARR> <USER SPECIALITY="D">1st year undergraduate student, female, 4 years of searching experience.</USER>

APPENDIX B. System evaluation results

In this section, evaluation results and the comparison of various types of retrieval systems for the Navigational retrieval task of NTCIR-4 using the current test collection are presented.

B-1 Summary of evaluated systems

For NTCIR-4, five groups and the authors submitted 86 run results from their search systems. We selected 10 typical runs shown below and compared their system behaviors mainly from the viewpoint of the types of information they use, i.e., whether they use anchor text information, link information but no anchor text information, or no link information or anchor text information but content information. The 10 selected runs are listed in Table A, in descending order of MRR values as described later. The evaluation results of all runs can be found in [10].

Characteristics of the systems corresponding to the 10 selected runs are as follows:

FR06: A system using site anchor text for indexing Web documents (see K3100-tt-02 in [1]).

FR025: A system using expanded anchor texts for indexing Web documents that are pointed to by them (run by the authors; see ORGREF-AT40-P1 in [1]).

FR061: A system using one-hop forward link analysis to expand retrieval sets retrieved by a content-based system with tf-idf ranking (run by the authors; see ORGREF-OT-D-LF2 in [1]).

FR077: A system using Okapi/BM25, pseudo-relevance feedback and PageRank [11].

FR067: A system using one-hop backward link analysis to expand retrieval sets retrieved by a content-based system with tf-idf ranking (run by the authors; see ORGREF-OT-DT-LB2 in [1]).

FR084: A system using the Relevance-based Superimposition Model with tf-idf ranking in combination with depth of URL hierarchy [12].

FR043: A Boolean-type system with ranking by tf-idf and weights on html tags (baseline run by the authors using Namazu [15]; see ORGREF-NMZ-AND in [1]).

FR081: An interactive system using Microsoft IIS Index Server (see W3SJP2003-001 in [1]).

FR076: A Boolean-type system with ranking by tf-idf and weights on html title and heading tags (run by the authors; see ORGREF-OT-DT in [1]).

FR082: A system using a probabilistic model based on gram-based indices of textual contents [13].

Table A. Characteristics of systems and evaluation results

RunID: Indicates the system run that generated the results.
TopicPart: Indicates the part of the topic used. The characters 'T', 'D', and 'B' indicate TITLE, DESC, and BACK respectively.
QExpan: Indicates if query expansion is used.
ContInfo: Indicates if full text content information is used for searching for or ranking documents.
LinkInfo: Indicates if link information is used for searching for or ranking documents.
AnchorInfo: Indicates if anchor text information is used for searching for or ranking documents.
MRR: Indicates mean reciprocal rank at top-100 document level.
DCG: Indicates discounted cumulative gain at top-100 document level.

RunID	TopicPart	Qexpan	ContInfo	LinkInfo	AnchorInfo	MRR	DCG
FR06	T	no	no	yes	Yes	0.4651	2.1335
FR025	T	no	no	yes	yes	0.4141	1.7468
FR061	T	no	yes	yes	no	0.2406	1.0911
FR077	T	yes	yes	yes	no	0.1773	1.0014
FR067	T	no	yes	yes	no	0.1164	0.8152
FR084	T	no	yes	no	no	0.0926	0.4543
FR043	T	no	yes	no	no	0.0920	0.6673
FR081	TDB	yes	yes	no	no	0.0872	0.5977
FR076	T	no	yes	no	no	0.0847	0.6666
FR082	T	yes	yes	no	no	0.0455	0.4201

B-2 Summary of evaluation results

We computed the effectiveness of individual run results shown in Section B-1 based on the evaluation method described in Section 7 on the DS-2 document data set at the relevance level of RL-1.

The evaluation results are shown in Table A, arranged in descending order of MRR.

The systems that utilize anchor text information, although their retrieval methods differ, have the highest scores. Besides these runs, several runs utilizing link information performed well. Runs using content information performed poorly. Although the results of only 10 runs are discussed here, the trend is in agreement with the evaluation results of all runs at NTCIR-4 [10].

Figure A shows graphs of cumulative numbers of topics for which relevant documents were retrieved by the 10 systems.

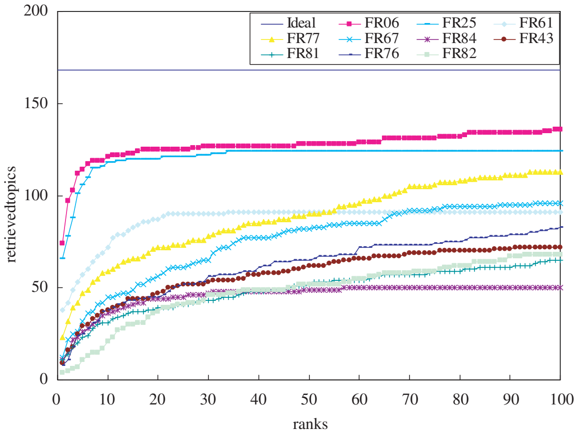


Fig. A Cumulative number of topics for which relevant documents were retrieved on (DS-2) and (RL-1).

We can see a tendency that curves of runs based on anchor information (FR06 and FR25) rise rapidly within rank 10 and are almost level thereafter, those of runs based on content information rise gradually over the entire rank range, while those of runs based on link information fall between the anchor-based runs and content-based runs. By inspecting curves for the last five of the 10 runs in Table A, the disagreement of the descending order of DCG value and that of MRR for the last five runs can be seen to occur because MRR favors the systems that retrieve the relevant documents at higher ranks.

APPENDIX C. Definitions of DCG and MRR

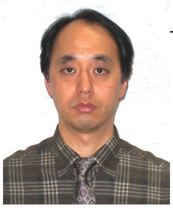
DCG, as described in [3] is an evaluation measure that takes account of multi-grade relevance and the ranking of the relevant documents, and is defined as follows.

$$dcg(i) = \begin{cases} g(i) & \text{if i=1} \\ dcg(i-1) + g(i)/\log_b(i) & \text{otherwise,} \end{cases}$$

$$g(i) = \begin{cases} a & \text{if d(i)} \in \text{A} \\ b & \text{if d(i)} \in \text{B} \\ c & \text{if d(i)} \in \text{C} \end{cases}$$

d(*i*) is the ith ranking of the document and A, B and C indicate the sets of highly relevant, fairly relevant and partially relevant documents respectively. The value of 2 was used for the base of logarithms, *b*. In this paper, (a, b, c) = (3, 2, 0) is used as the weight. A and B correspond to “relevant” and “Partially relevant” documents respectively and C = ϕ.

MRR, as described in [3], is the average of the reciprocal of the highest ranking of the relevant document over all topics.



Keizo OYAMA

Keizo OYAMA received the B.E., M.E. and Dr. Eng. from the University of Tokyo in 1980, 1982 and 1985, respectively. He is a professor at National Institute of Informatics (NII), Japan. He is also an adjunct professor at the Graduate University for Advanced Studies (Sokendai). His research interests are structured text processing, web information retrieval and utilization, and full-text search technologies. He is a member of IEICE, IPSJ, JSIMS and DBSJ.

Haruko ISHIKAWA

Haruko ISHIKAWA graduated from Department of Physics at Gakushuin University, Japan in 1992, awarded PhD from the department of Mechanical engineering at the University of Queensland in Australia in 2000. She has been a contract researcher at CSIRO Exploration & Mining in Australia, a postdoctor at Pierre & Marie Curie University in France and currently a post doctor at National Institute of Informatics in Japan. Her current research interests include web content analysis and their evaluation methodologies.

Koji EGUCHI

Koji EGUCHI received the B.E from Doshisha University, Japan, in 1993, and the M.E. and Dr.Eng. from Kansai University, Japan, in 1995 and 1999, respectively. From 1999 to 2000, he was a Research Associate at National Center for Science Information Systems, Japan. In 2000, he joined National Institute of Informatics, Japan, where, since 2004, he has been an Assistant Professor. Since 2002, he has held a concurrent position at the Graduate University for Advanced Studies, Japan. His current research interests include information retrieval, web content analysis, and their evaluation methodologies. He is a member of ACM, IPSJ, IEICE and JSAI.

Akiko AIZAWA

Akiko AIZAWA graduated from the Department of Electronics at the University of Tokyo in 1985 and completed her doctoral studies in electrical engineering in 1990. She was a visiting researcher at the University of Illinois at Urbana-Champaign from 1990 to 1992. At present, she is an professor at National Institute of Informatics and the Graduate University for Advanced Studies (Sokendai). Her research interests include statistical text processing, linguistic resources construction, and corpus-based knowledge acquisition.