An Evaluation of the Web Retrieval Task at the Third NTCIR Workshop

Koji Eguchi † **Keizo Oyama** † **Emi Ishida** ††
Noriko Kando † **Kazuko Kuriyama** †††

† **National Institute of Informatics**
{*eguchi,oyama,kando*}*@nii.ac.jp*
†† **Surugadai University**
emi@surugadai.ac.jp
††† **Shirayuri College**
kuriyama@shirayuri.ac.jp

1 Introduction

We have investigated the evaluation methods for measuring retrieval effectiveness of Web search engine systems, attempting to make them suitable for real Web environment. With this objective, we conducted 'Web Retrieval Task' at the Third NTCIR Workshop ('NTCIR-3 WEB') from 2001 to 2002 [1, 2, 3]. Using this NTCIR-3 WEB, we built a re-usable test collection that is suitable for evaluating Web search engine systems, and evaluated the retrieval effectiveness of a certain number of Web search engine systems.

TREC Web Tracks [4] are well-known workshops that have an objective to research the retrieval of large-scale Web document data. Past TREC Web Tracks have used data sets extracted from 'the Internet Archive'[1]or pages gathered from the '.gov' domain as document sets. They assessed the relevance only on information given in English. NTCIR-3 WEB was another workshop that has used 100-gigabyte and/or 10-gigabyte document data that were mainly gathered from the '.jp' domain. Relevance judgment was performed on the retrieved documents that are written in Japanese or English, partially considering hyperlinks. By considering the hyperlinks, a 'hub page' that gives out-links to multiple 'authority pages' [5] may be judged as relevant even if these do not include sufficient relevant information in them. 16 groups enrolled to participate in the NTCIR-3 WEB, and seven of these groups submitted run results.

2 Task Description

The NTCIR-3 WEB was composed of the following tasks for the two document data sets with the sizes of: (I) 100 gigabytes, and (II) 10 gigabytes, respectively.

- **A:** Survey Retrieval Tasks (Topic Retrieval Task and Similarity Retrieval Task)
- **B:** Target Retrieval Task
- **C:** Optional Tasks (Search Results Classification Task and Speech-Driven Retrieval Task)

The Survey Retrieval Tasks assumed the user model where the user attempts to comprehensively find documents relevant to his/her information needs. Three types of queries were supposed: query term(s) and sentence as 'Topic Retrieval Task', and query document(s) as 'Similarity Retrieval Task'. We describe the details of the Topic Retrieval Task below, but omit the details of the Similarity Retrieval Task [1, 2], of which results submission was less than we had expected. The Topic Retrieval Task is similar to a traditional

[1] ⟨http://www.archive.org/⟩

Table 1: Fundamental statistics of the document sets

Statistics of NW100G-01	
(1-1) # of crawled sites *	97,561
(1-2) max. # of pages within a site	1,300
(1-3) # of crawled pages **	11,038,720
(1-4) # of pages for searching	15,364,404
(1-5) # of links connected from (1-3)	78,175,556
(1-6) # of links connected from (1-3) to (1-4) ***	64,365,554
Statistics of NW10G-01	
(2-1) # of crawled sites *	97,561
(2-2) max. # of pages within a site	20
(2-3) # of crawled pages **	1,445,466
(2-4) # of pages for searching	4,849,714
(2-5) # of links connected from (2-3)	11,642,167
(2-6) # of links connected from (2-3) to (2-4) ***	9,885,538

(*) Aliased sites are not included.
(**) *i.e.*, # of pages included in the document data for providing. Aliased sites are not included.
(***) *i.e.*, # of pages included in the document data for reference. The existence of the other pages, *i.e.*, (1-6)-(1-5) or (2-6)-(2-5), could not be confirmed.

Table 2: Proportion of the page numbers for each language

Language	Proportion
Japanese	90. %
English	8.3 %
Simp. Chinese	0.05%
Korean	0.03%
Trad. Chinese	0.02%
West Europian	0.01%
Other Languages *	0.01%
No Text Content	0.78%
Not Idenfied	0.02%

(*) Russian, East Europian, Thai, Hebrew, Arabic, and Turkish

ad-hoc retrieval as in TRECs or NTCIR Workshops, and so ensures the reusability of the test collection. The participants in the Topic Retrieval Task had to submit at least two lists of their run results: that of the run using only the topic field of ⟨TITLE⟩ and that of the run using only ⟨DESC⟩. They could optionally submit their run results using other topic fields. The details of the topic formats are described in Sect. **3.2**.

The Target Retrieval Task aimed to evaluate the effectiveness of the retrieval, by supposing a user model where the user requires just one answer, or only a few answers. The precision of the ranked search results was emphasized in this study. The runs were evaluated using the 10 top-ranked documents retrieved for each topic. Several evaluation measures were applied, as is described in Sect. **4**.

Two optional tasks were adopted: (i) 'Search Results Classification Task' that tried to evaluate classification-based output presentation, and (ii) 'Speech-Driven Retrieval Task' that evaluated searches driven by spoken queries against Web documents. We have omitted describing the details of these tasks in this paper, but overviews can be seen in References [1, 6].

3 Web Test Collection Building

The 'Web Test Collection' is composed of: (i) the document sets, (ii) the topics, and (iii) the list of relevance judgment results for each topic. Each of these components is suitable for the real Web environment, as is described in Sects. **3.1**, **3.2**, and **3.4**, respectively. Moreover, pooling had to be performed before the relevance judgments, as described in Sect. **3.3**.

3.1 Document Sets

The document sets are explicitly specified for the test collections. In the NTCIR-3 WEB, we prepared two sets of document data gathered from the '.jp' domain: (i) document data over 100 gigabytes ('NW100G-01'), and (ii) 10-gigabyte document data ('NW10G-01'). We also provided two separate lists of documents that were connected from the individual documents included in the NW100G-01 and NW10G-01 data, but not limited to the '.jp' domain. These four data sets were used for searching in the NTCIR-3 WEB.

Fundamental statistics of the document sets are shown in **Table 1**. In the NW100G-01, the proportion of the page numbers for each language is shown in **Table 2**. The figures are rough estimates based on

character sets specified in content-type field[2]. Undetected English pages may be included in other language classes. The crawling strategy is described in Reference [1, 2]. The participants were allowed to process the NW100G-01 and NW10G-01 data only within the 'Open Laboratory' located at the National Institute of Informatics (NII) in Japan[3].

3.2 Topics

The organizers provided 'topics' that are statements of information needs rather than queries. The topic format was basically inherited from previous TRECs and NTCIR Workshops, but some tags were redefined for the NTCIR-3 WEB, such as ⟨TITLE⟩ and ⟨NARR⟩ tags[4]. Those tags having the following meanings specify each field.

- ⟨TITLE⟩ provides up to three terms that were specified by the topic creator, simulating the query terms in real Web search engines. The topic creator selected one of the three search strategies –AND, OR or the combination of them–, deemed suitable for obtaining the needed information using search engines, and specified up to three terms for inputting into the search engine. The terms in the ⟨TITLE⟩ tag are listed in the order of importance for searching.
- ⟨NARR⟩ ('narrative') describes, in a few paragraphs, the background to the purpose of the retrieval, the term definitions, and the relevance judgment criteria. These are flanked by ⟨BACK⟩, ⟨TERM⟩, and ⟨RELE⟩ tags, respectively, in ⟨NARR⟩. Any of them may be omitted.

We also describe the definition of ⟨DESC⟩, the same as TRECs or NTCIR Workshops as follows, but omit describing the definitions of other tags.

- ⟨DESC⟩ ('description') represents the most fundamental description of the user's information needs in a single sentence.

All of the above topics are written in Japanese. The usable fields and mandatory fields varied according to the tasks [1, 2].

3.3 Pooling

All the topics are created without using any search systems or any relevance assessment[5]. Therefore, some of the topics were not suitable for use in a comparison of retrieval effectiveness. Consequently, we applied several topic selection strategies, such as the 'shallow pooling' described below, to discard any inappropriate topics, such as ones that had few relevant documents [1, 2]. As a result, we used 47 topics for evaluation.

'Shallow pooling' is a sampling method that takes the 20 highest-ranked documents from each run result submitted by a participant, ranking them in the order of a meta-search-engine strategy. By assessing the relevance of each document included in the 'shallow pool,' we were able to discard some of the topics.

Using the topics selected by the shallow pooling, we performed 'deep pooling,' which took the potentially large number of top-ranked documents from each run result and merged them, as in the pooling methods previously used in TRECs or NTCIR Workshops. Through the pooling stage, we obtained a subset of the document data, called the 'pool', which was used to estimate the relevant documents included in the document data especially for the evaluation of the Survey Retrieval Task. In the pooling task, we took the 100 top-ranked documents from each run result. Moreover, we performed ranking on the pooled documents in the order of the meta-search-engine strategy, using the same process as in the shallow pooling stage.

[2]The procedure to estimate the language proportion is described in Reference [2].

[3]Participants were allowed to take out the processed data such as index files, but not the original document data. To perform the data processing, remote access to the individual host computers in the Open Laboratory was allowed.

[4]We also added some other tags as described in Reference [1, 2].

[5]By way of exception, the ⟨RDOC⟩ field [1] were created using relevance assessment results using a baseline search system.

3.4 Relevance Assessment

Pooled documents that were composed of the top-ranked search results submitted by each participant were considered to be the relevant document candidates. The assessors judged the relevance of the pooled documents only on the basis of the information given in Japanese or in English, although some of the documents included in the document data were written in languages other than Japanese or English. The assessors judged the 'Multi-Grade Relevance' of the individual pooled documents as: highly relevant, fairly relevant, partially relevant, or irrelevant.

By the way, Web pages are represented in various ways, so that in one example, an 'information unit' on the Web could be hyperlinked pages, while in another, it could be an individual page, or a passage included in a page. Assuming an information unit on the Web to be a page, a 'hub page' that gives out-links to multiple 'authority pages' [5] must be judged as irrelevant if these do not include sufficient relevant information in themselves. However, in the Web environment, this type of hub pages are sometimes more useful for the users than the relevant pages defined by the assumption.

The NTCIR-3 WEB attempted to incorporate concepts of two additional information units into the relevance assessment, assuming that a group of hyperlinked pages or a passage can be an information unit, so we defined the following three document models:

One-click-distance document model This is where the assessor judges the relevance of a page when he/she browses the page and its out-linked pages that are included in the pool, but not all of the out-linked pages, assuming that most of the relevant documents were included in the pool.

Page-unit document model This is where the assessor judges the relevance of a page only on the basis of the entire information given by itself, as is performed conventionally.

Passage-unit document model This is where the assessor specifies the passages that provides evidence of relevance, which he/she uses to judge the pages relevant.

4 Evaluation Measures

4.1 Precision/Recall and Discounted Cumulative Gain

In evaluating the run results of each participant's search engine system, we focused on up to 1,000 top-ranked documents for the Survey Retrieval Tasks, and up to 10 top-ranked documents for the Target Retrieval Task.

For the Survey Retrieval Tasks, we applied the two types of evaluation measures: (i) those based on precision and/or recall [6], and (ii) those with discounted cumulative gain ('DCG') [7]. For the Target Retrieval Task, we applied the three types of measure: the aforementioned measures in (i) and (ii), and a weighted reciprocal rank measure (iii), which will be described in Sect.**4.2**.

Although the one-click-distance document model was partly applied in the relevance assessment, as described in Sect. **3.4**, almost all the evaluation measures were designed by assuming a page to be the basic unit. However, for a given relevant document set, an important factor is the differences between the two document models: the one-click-distance document model, and the page-unit document model. In computing the values of the evaluation measures for each run result, we used two types of relevant document sets, according to which of the two document models was used.

4.2 Weighted Reciprocal Rank

The 'Mean Reciprocal Rank' measure [8] ('MRR') is often used in evaluating question answering systems, and is defined as the average over all the questions of the reciprocal of the rank of the first appearing answer for each question. We applied the idea of the MRR to evaluate the run results of the Target Retrieval Task, so that we made it generalized for multi-grade relevance. In the NTCIR-3 WEB, we proposed a new measure,

[6] Almost those evaluation measures can be computed using 'trec_eval', a program that evaluates TREC results. This is available at ⟨ftp://ftp.cs.cornell.edu/pub/smart/trec_eval. v3beta.shar⟩.

Table 3: System ranking of Survey and Target Retrieval Tasks

Topic part used	Survey, Topic Retrieval Task				Target Retrieval Task			
	aprec	rprec	dcg(100)	dcg(1K)	prec(10)	dcg(10)	wrr(10)	%nf(10)
TITLE	GRACE-LA1-1	**GRACE-LA1-2**	GRACE-LA1-1	GRACE-LA1-1	GRACE-LB-1	GRACE-LB-1	**K3100-13**	**K3100-13**
TITLE	**GRACE-LA1-2**	GRACE-LA1-1	**GRACE-LA1-2**	**GRACE-LA1-2**	**GRACE-LB-2**	**GRACE-LB-2**	GRACE-LB-1	**K3100-14**
TITLE	OKSAT-F-04	ORGREF-LA1-6	OKSAT-F-04	OKSAT-F-04	**K3100-14**	**K3100-14**	**GRACE-LB-2**	**GRACE-LB-2**
TITLE	ORGREF-LA1-6	OKSAT-F-04	ORGREF-LA1-6	ORGREF-LA1-6	**K3100-13**	**K3100-13**	**K3100-14**	GRACE-LB-1
TITLE	**K3100-05**	**K3100-05**	**K3100-05**	**K3100-05**	ORGREF-LB-6	ORGREF-LB-6	ORGREF-LB-6	ORGREF-LB-6
TITLE	**K3100-06**	**K3100-06**	**K3100-06**	**K3100-06**	UAIFI8	NAICR-I-B-1	ORGREF-LB-3	UAIFI8
TITLE	NAICR-I-A1-4	NAICR-I-A1-4	NAICR-I-A1-4	UAIFI4	UAIFI7	UAIFI8	NAICR-I-B-1	UAIFI7
TITLE	ORGREF-LA1-5	UAIFI3	UAIFI3	UAIFI3	NAICR-I-B-1	ORGREF-LB-5	ORGREF-LB-5	NAICR-I-B-1
TITLE	UAIFI3	ORGREF-LA1-5	UAIFI4	NAICR-I-A1-4	ORGREF-LB-5	UAIFI7	UAIFI8	ORGREF-LB-3
TITLE	UAIFI4	UAIFI4	ORGREF-LA1-5	ORGREF-LA1-5	ORGREF-LB-3	ORGREF-LB-3	ORGREF-LB-4	ORGREF-LB-5
TITLE	ORGREF-LA1-3	ORGREF-LA1-3	ORGREF-LA1-3	ORGREF-LA1-3	ORGREF-LB-4	ORGREF-LB-4	UAIFI7	ORGREF-LB-4
TITLE	ORGREF-LA1-1	ORGREF-LA1-1	ORGREF-LA1-1	ORGREF-LA1-1	ORGREF-LB-1	ORGREF-LB-1	ORGREF-LB-1	ORGREF-LB-1
TITLE	ORGREF-LA1-4	ORGREF-LA1-4	ORGREF-LA1-4	ORGREF-LA1-4	ORGREF-LB-2	ORGREF-LB-2	ORGREF-LB-2	ORGREF-LB-2
TITLE	ORGREF-LA1-2	ORGREF-LA1-2	ORGREF-LA1-2	ORGREF-LA1-2				
DESC	**GRACE-LA1-4**	**GRACE-LA1-4**	**GRACE-LA1-4**	GRACE-LA1-3	**GRACE-LB-4**	GRACE-LB-3	**GRACE-LB-4**	**GRACE-LB-4**
DESC	GRACE-LA1-3	GRACE-LA1-3	GRACE-LA1-3	**GRACE-LA1-4**	GRACE-LB-3	**GRACE-LB-4**	GRACE-LB-3	NAICR-I-B-2
DESC	OKSAT-F-06	OKSAT-F-06	OKSAT-F-06	OKSAT-F-06	UAIFI5	UAIFI6	**K3100-15**	GRACE-LB-3
DESC	NAICR-I-A1-3	NAICR-I-A1-3	NAICR-I-A1-3	NAICR-I-A1-3	UAIFI6	UAIFI5	NAICR-I-B-2	UAIFI5
DESC	**K3100-07**	UAIFI2	UAIFI1	UAIFI2	NAICR-I-B-2	**K3100-15**	UAIFI6	**K3100-15**
DESC	UAIFI1	UAIFI1	UAIFI2	UAIFI1	**K3100-15**	NAICR-I-B-2	**K3100-16**	UAIFI6
DESC	UAIFI2	**K3100-07**	**K3100-07**	**K3100-08**	NAICR-I-B-3	NAICR-I-B-3	UAIFI5	**K3100-16**
DESC	NAICR-I-A1-2	**K3100-08**	**K3100-08**	**K3100-07**	NAICR-I-B-4	NAICR-I-B-4	NAICR-I-B-3	NAICR-I-B-3
DESC	**K3100-08**	NAICR-I-A1-2	NAICR-I-A1-2	NAICR-I-A1-2	**K3100-16**	**K3100-16**	NAICR-I-B-4	NAICR-I-B-4
DESC	NAICR-I-A1-1	NAICR-I-A1-1	NAICR-I-A1-1	NAICR-I-A1-1				

'aprec' indicates the average precision (non-interpolated).

'rprec' indicates the R-precision, *i.e.*, the average of the precision after $|R|$ documents were retrieved, where $|R|$ indicates the number of relevant documents for each topic.

'prec(10)' indicates the precision at the 10-document level.

'dcg(1K)', 'dcg(100)' and 'dcg(10)' indicate the DCG values at the 1,000-, 100- and 10-document level, respectively.

'wrr(10)' indicates the WRR value at the 10-document level.

'%nf(10)' indicates the percentage of topics for which no relevant documents were retrieved at the 10-document level.

the 'Weighted Reciprocal Rank' ('WRR') as the mean value of the wrr(m), defined by the following equations over all the topics [1, 2]:

$$\mathrm{wrr}(m) = \max\left(\mathrm{r}(m)\right) \ , \tag{1}$$

$$\mathrm{r}(m) = \begin{cases} \delta_h / \left(i - 1/\beta_h\right) & \text{if } (\mathrm{d}(i) \in H \wedge 1 \leq i \leq m) \\ \delta_a / \left(i - 1/\beta_a\right) & \text{if } (\mathrm{d}(i) \in A \wedge 1 \leq i \leq m) \\ \delta_b / \left(i - 1/\beta_b\right) & \text{if } (\mathrm{d}(i) \in B \wedge 1 \leq i \leq m) \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

where m indicates the rank at the cut-off level in the run results, and the weight coefficients satisfy $\delta_h \in \{1, 0\}$, $\delta_a \in \{1, 0\}$, $\delta_b \in \{1, 0\}$, and $\beta_b \geq \beta_a \geq \beta_h > 1$, respectively. In Equation (2), the term $(-1/\beta_x)$, $x \in \{h, a, b\}$ can be omitted when the value of β_x is sufficiently large.

5 Evaluation Results

5.1 Summary of Participation

Six groups[7]submitted their completed run results, with the organizers also submitting the results from their own search system along with those of the participants in an attempt to try to improve the comprehensiveness of the pool. The individual participating groups pursued various objectives, such as using a re-ranking method with link analysis, using an anchor-based retrieval method, using a distributed information retrieval method, using a pseudo-relevance method, and using a gram-based indexing method[8]. Unfortunately, only two of the participating groups could submit the run results for the Similarity Retrieval Task.

[7]The participated groups are (1) Nara Institute of Science and Technology and Communication Research Laboratory, (2) NEC Corporation, (3) Osaka Kyoiku University, (4) University of Aizu, (5) University of Library and Information Science and National Institute of Advanced Industrial Science and Technology, and (6) University of Tokyo and RICOH Co. Ltd.

[8]Papers on the details will be available online at ⟨http://research.nii.ac.jp/ntcir/publication1-en.html⟩.

5.2 Experimental Conditions

In evaluating the run results against 100-gigabyte and 10-gigabyte data, we used combinations of $\{PL_1, PL_2\} \times \{DM_1, DM_2\} \times \{RL_1, RL_2\}$, which were defined as follows:

Pooling Methods

- **(PL_1) Pooling for large-scale runs** The list of relevant documents, with relevance judged on individual documents in the pools. They were taken from the run results against the 100-gigabyte and 10-gigabyte data sets.
- **(PL_2) Pooling for small-scale runs** The list of relevant documents, with relevance judged on individual documents in the pools. They were taken from the run results against 10-gigabyte data set.

Document Models[9] (as described in Sect. **3.4**)

- **(DM_1) One-click-distance document model**
- **(DM_2) Page-unit document model**

Relevance Levels

- **(RL_1) Rigid relevance level** When using evaluation measures based on precision and/or recall, we considered the document to be relevant if it was highly relevant or fairly relevant, and otherwise considered it to be irrelevant.
- **(RL_2) Relaxed relevance level** When using evaluation measures based on precision and/or recall, we considered the document to be relevant if it was highly relevant, fairly relevant, or partially relevant. Otherwise, we considered it to be irrelevant.

5.3 Discussion on Effects of Hyperlinks

We computed the effectiveness of individual run results as shown in Sect. **5.1** using the respective evaluation measures described in Sect. **4** and using the conditions as described in Sect. **5.2**. For the Survey, Topic Retrieval Task and the Target Retrieval Task against 100-gigabyte data, we ranked the run results in the orders of several evaluation measures using the one-click-distance document model (DM_1) and the rigid relevance level (RL_1), as shown in **Table 3**.

Focusing on the Target Retrieval Task (the right part of the table), we observe the distribution of run IDs that were carried out by the systems based on not only page content but also hyperlink information (underlined run ID codes). As the results, it suggests that the link-based systems perform more effectively with short queries such as the TITLEs than longer queries such as the DESCs. Moreover, focusing on the TITLE-only runs in both tasks (the upper part of the table), we compared the distribution of underlined run ID codes. As the results, it suggests that the link-based systems using short queries perform more effectively for highly ranked documents such as in the Target Retrieval Task than for entire ranked results such as in the Survey Retrieval Task[10].

6 Ongoing NTCIR-4 WEB Task

The WEB Task at the 4th NTCIR Workshop (NTCIR-4 WEB) attempts to push ahead researches of information access systems for large-scale Web documents, making use of the experiences of the NTCIR-3 WEB. The organizers investigated actual use of the Web from various viewpoints, and designed the following tasks to evaluate the required fundamental techniques.

- **A:** Informational Retrieval Task
- **B:** Navigational Retrieval Task

[10] 'GRACE-LA1-1' and 'GRACE-LB-1' did not use link information but were highly ranked, however, they can be considered as exceptions since their system parameters were different from those of the same group's other run ID codes that start with 'GRACE'.

C: Geographic Information Task[11]
D: Topical Classificaiton Task

The names of the task A and B were derived from Broder's taxonomy [9], as TREC Web Tracks were. The task A was designed to evaluate effectiveness of the search engines from the viewpoint of topic relevance, by combining the Survey Retrieval Task and the Target Retrieval Task at NTCIR-3 WEB. The task A was further emphasized on evaluation considering hyperlink relationship and content duplication[12]. The task B was designed to evaluate effectiveness of the search engines, assuming that a user is motivated to find a small number of typical Web pages of a certain object, such as a person, shop, restaurant or facility. The task C investigates the feasibility to evaluate techniques that extract geographical descriptions from the Web pages relevant to a given viewpoint. The task D attempts to evaluate techniques for supporting user's browsing process by means of classification-based output presentation when the user submits very short queries that have ambiguity[13]. The workshop meeting will be held in June 2004 at National Institute of Informatics (NII) in Tokyo, Japan. You can see the up-to-date information on NTCIR-4 WEB on the Web site [3].

Acknowledgements

This work was partially supported by the Grants-in-Aid for Scientific Research on Priority Areas of "Informatics" (#13224087) and for Encouragement of Young Scientists (#14780339) from the Ministry of Education, Culture, Sports, Science and Technology, Japan. We greatly appreciate the efforts of all the participants of the Web Retrieval Task at the Third NTCIR Workshop. We also appreciate the useful advice of the Web Retrieval Task Advisory Committee, and Professor Jun Adachi, National Institute of Informatics.

References

[1] K. Eguchi, K. Oyama, E. Ishida, N. Kando and K. Kuriyama: "Overview of the Web Retrieval Task at the Third NTCIR Workshop", Proc. of the 3rd NTCIR Workshop on Research in Information Retrieval, Automatic Text Summarization and Question Answering (2003).

[2] K. Eguchi, K. Oyama, E. Ishida, N. Kando and K. Kuriyama: "Evaluation Methods for Web Retrieval Tasks Considering Hyperlink Structure", IEICE Transactions on Information and Systems, Vol.E86-D, No.9, pp.1804-1813 (2003).

[3] "NTCIR-WEB", ⟨http://research.nii.ac.jp/ntcweb/⟩.

[4] "TREC Web Track", ⟨http://es.csiro.au/ TRECWeb/⟩.

[5] J. Kleinberg: "Authoritative sources in a hyperlinked environment", Proc. of the 9th ACM SIAM Symposium on Discrete Algorithms (1998).

[6] A. Fujii and K. Itou: "Evaluating speech-driven IR in the NTCIR-3 Web Retrieval Task", Proc. of the 3rd NTCIR Workshop on Research in Information Retrieval, Automatic Text Summarization and Question Answering (2003).

[7] K. Järvelin and J. Kekäläinen: "IR evaluation methods for retrieving highly relevant documents", Proc. of the 23rd Annual International ACM SIGIR Conference, pp. 41–48 (2000).

[8] E. Voorhees: "The TREC-8 Question Answering Track report", Proc. of the 8th Text REtrieval Conference, NIST Special Publication 500-246, pp. 77–82 (1999).

[9] A. Broder: "A taxonomy of web search", ACM SIGIR Forum, Vol.36, No.2, pp. 3-10 (2002).

[11] This task is organized by Masatoshi Arikawa and Takeshi Sagara at the University of Tokyo.

[12] For a part of the duplicate documents, we treat them as irrelevant or partially relevant in the evaluation although they are judged as relevant by an assessor. Consequently, run results that contained the duplicated documents will be expected to pay a penalty.

[13] The task D is based on the 'Search Results Classification Task' at the NTCIR-3 WEB, which was proposed as a pilot study and adopted as one of the 'Optional Tasks' at the NTCIR-3 WEB, but no classification results were submitted on time.