Overview of the Informational Retrieval Task at NTCIR-4 WEB

Koji Eguchi † Keizo Oyama † Akiko Aizawa † Haruko Ishikawa †

† National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan
{eguchi, oyama, akiko, haruko}@nii.ac.jp

Abstract

This paper gives an overview of the Informational Retrieval Task 2 that was conducted from 2003 to 2004 as a subtask of the WEB Task at the Fourth NTCIR Workshop ('NTCIR-4 WEB'). In the Informational Retrieval Task, we attempted to assess the retrieval effectiveness of Web search engine systems from a viewpoint of topical relevance, and to build a re-usable test collection suitable for evaluating Web search engine systems from such a viewpoint. We used a 100-gigabyte document dataset that was mainly gathered from the '.jp' domain. Relevance judgments were performed on the retrieved documents written in Japanese or English, partially considering the relationshiop between the pages referenced by hyper-links. We also investigated the evaluation methods considering non-redundancy of contents and diversity of queries.

Keywords: *Web Information Retrieval, Evaluation Methods, Test Collections.*

1 Introduction

This paper gives an overview of the Informational Retrieval Task 2 that was conducted from 2003 to 2004 as a subtask of the WEB Task at the Fourth NTCIR Workshop ('NTCIR-4 WEB'). In the Informational Retrieval Task, we attempted to assess the retrieval effectiveness of Web search engine systems from a viewpoint of topical relevance, and to build a re-usable test collection suitable for evaluating Web search engine systems from such a viewpoint. The name of the task was derived from Broder's taxonomy [1].

The Informational Retrieval Task is similar to a traditional ad-hoc retrieval [10, 7] at the point of focusing on the topical relevance. However, this subtask is different from these at the following points: (1) The relationshiop between the pages referenced by hyper-links were considered in relevance judgments; (2) Non-redundancy of page contents were taken into account in evaluation measures.

The task design is also similar to the Topic Distillation Task in Web Track at TREC 2002 and TREC 2003 [2, 3] at the point of consideration of hyper-links in relevance assessments. However, the methods of relevance assessment considering hyper-links are slightly different from the one of the Informational Retrieval Task and the one of the Topic Distillation Task.

The Informational Retrieval Task is derived from the 'Survey Retrieval Task' and the 'Target Retrieval Task' conducted in the Web Retrieval Task at the Third NTCIR Workshop ('NTCIR-3 WEB'), and is further emphasized on the consideration of hyper-links and non-redundancy of contents as mentioned above.

We used the 100-gigabyte document dataset ('NW100G-01') that was constructed at the Third NTCIR Workshop. Those were mainly gathered from the '.jp' domain. We also created the topics —*i.e.*, statements of information needs—, considering diversity of query expressions, such as the query expressed by a single term having a vague or broader meaning, and query expressions specified by several persons for the same information needs.

2 Task Description

The Informational Retrieval Task assumed two user models: (i) the model where the user attempted to comprehensively find documents relevant to his/her information needs, and (ii) the model where the user requires just one or only a few relevant documents at the highly ranked documents.

Two types of queries were supposed: (i) query term(s) specified in topic fields of ⟨TITLE⟩, ⟨ALT0⟩, ⟨ALT1⟩, ⟨ALT2⟩ and ⟨ALT3⟩, and (ii) sentence(s) specified in a topic field of ⟨DESC⟩. The participants had to submit at least six lists of their run results: that of the run using each of the topic fields metioned above. The details of the topic formats are described in Section **3.2.1**.

The participating groups submitted their run results using the identification numbers of 1,000 retrieved

documents ranked for each topic[1] . The run results of both 'automatic' and 'interactive' systems were accepted. Any search systems involving manual intervention during the search process were deemed 'interactive,' with all the others being 'automatic'.

The participating groups were requested to report which fields of the topics were used in the automatic or interactive systems. In evaluating the systems, comparisons of their effectiveness should be performed separately, according to which runs are 'automatic' or 'interactive,' and which fields of the topic are used.

3 The Web Test Collection

The 'Web Test Collection' was composed of the followings:

- the document set,
- the topics, and
- the list of relevance judgment results for each topic.

Each of these components was designed to be suitable for the real Web environment, as is described in Sections **3.1**, **3.2** and **3.4**, respectively. Moreover, pooling has to be performed before relevance judgments, as described in Section **3.3**.

3.1 Document Set

The document sets are explicitly specified for the test collections. In the NTCIR-4 WEB, we used 'NW100G-01' dataset that was constructed at the NTCIR-3 WEB as the document set. The NW100G-01 is composed of the document data gathered from the '.jp' domain. We also provided a separate list of documents that were connected from the individual documents included in the NW100G-01 dataset, but not limited to the '.jp' domain. These two datasets were used for processing at the NTCIR-4 WEB.

We stored the NW100G-01 dataset in a hard disk drive and delivered it to each participating group. In addition, for the purpose of handling the NW100G-01 dataset, the computer resources at the 'Open Laboratory' located at National Institute of Informatics were available only for the participants who request to use them.

3.2 Topics

3.2.1 Topic Format

The organizers provided 'topics' that were statements of information needs, and that also included typical query expressions.

The topic format was basically inherited from the NTCIR-3 WEB [5, 4], except for adding ⟨ALT0⟩, ⟨ALT1⟩, ⟨ALT2⟩ and ⟨ALT3⟩ and removing ⟨RDOC⟩ and ⟨CONC⟩. A pair of tags having the following meanings flanked each field:

- ⟨TOPIC⟩ specified the boundary of a topic.
- ⟨NUM⟩ indicated the topic identification number.
- ⟨TITLE⟩ gives 1-3 terms that are simulated by the topic creator to be similar to query terms used in real Web search engines. The terms in the ⟨TITLE⟩ are listed in their order of importance for searching. The ⟨TITLE⟩ has the attribute of 'CASE' that indicates the types of search strategies, as follows:
  (a) All of the terms have the relation one another that can be used as OR operator.
  (b) All of the terms have the relation one another that can be used as AND operator.
  (c) Only two of the terms have the relation that can be used as OR operator, and are specified by the attribute of 'RELAT'.
- ⟨DESC⟩ ('description') represented the most fundamental description of the user's information needs in a single sentence.
- ⟨NARR⟩ ('narrative') described, in a few paragraphs, the background to the purpose of the retrieval, the term definitions, and the relevance judgment criteria. These were flanked by ⟨BACK⟩, ⟨TERM⟩, and ⟨RELE⟩ tags, respectively, in ⟨NARR⟩. It was possible to omit some terms.
- ⟨ALT0⟩ ('alternative query 0') was created as the result of extracting the first appeared term in the ⟨TITLE⟩ field of the topic. The ⟨ALT0⟩ field has the term judged as being most important for searching, since the terms in the ⟨TITLE⟩ field were listed in the order of importance for searching. This tag was omitted when the ⟨TITLE⟩ field originally has only one term.
- ⟨ALT1⟩, ⟨ALT2⟩ and ⟨ALT3⟩ were created by three persons who were different than the topic creator, when he/she browsed the topic statement where the ⟨TITLE⟩ was deleted in advance. The format of these tags is the same as the one of ⟨TITLE⟩ except for the tag name. We omitted describing each tag when it was exactly the same —including the order of the topic terms— as the originally defined ⟨TITLE⟩. Therefore, any of those three tags may be omitted.
- ⟨USER⟩ ('user attributes') provided the attributes of the topic creator, *i.e.*, job title, gender, search experience, level of search skill, and level of familiarity with the topic.

All of the above topics were written in Japanese. A topic example and its English translation are shown in **Figure 1**.

[1] The participating groups also submitted a concise description of each run as well as the run results, as being described in **Appendix**.

⟨TOPIC⟩
⟨NUM⟩0001⟨/NUM⟩
⟨TITLE CASE=”c” RELAT=”2-3”⟩ オフサイド, サッカー, ルール ⟨/TITLE⟩
⟨DESC⟩ サッカーのオフサイドというルールについて説明されている文書を探したい ⟨/DESC⟩
⟨NARR⟩⟨BACK⟩ サッカーでオフサイドとはどういうルールなのかを知りたい。⟨/BACK⟩⟨TERM⟩ オフサイドはオフェンス側の反則である。オフサイドが適用される状況にはいくつかのパターンがあり、サッカーのルールの中で最もわかりにくいものである。⟨/TERM⟩⟨RELE⟩ 適合文書はオフサイドが適用される状況を説明しているもの ⟨/RELE⟩⟨/NARR⟩
⟨ALT0 CASE=”b”⟩ オフサイド ⟨/ALT0⟩
⟨ALT1 CASE=”b”⟩ オフサイド, 選手, 位置 ⟨/ALT1⟩
⟨ALT2 CASE=”b”⟩ オフサイド, サッカー ⟨/ALT2⟩
⟨ALT3 CASE=”b”⟩ サッカー, オフサイド, ルール ⟨/ALT3⟩
⟨USER⟩ 大学 2 年, 男性, 検索歴 4 年, 熟練度 3, 精通度 5⟨/USER⟩
⟨/TOPIC⟩

(a) An original sample topic

⟨TOPIC⟩
⟨NUM⟩0001⟨/NUM⟩
⟨TITLE CASE=”c” RELAT=”2-3”⟩offside, soccer, rule⟨/TITLE⟩
⟨DESC⟩ I want to find documents that explain the offside rule in soccer. ⟨/DESC⟩
⟨NARR⟩⟨BACK⟩ I want to know about the offside rule in soccer. ⟨/BACK⟩⟨TERM⟩ Offside is a foul committed by a member of the offense side. There are several patterns for situations in which the offside rule can be applied, and it is the most difficult soccer rule to understand. ⟨/TERM⟩⟨RELE⟩ Relevant documents must explain situations where the offside rule applies. ⟨/RELE⟩⟨/NARR⟩
⟨ALT0 CASE=”b”⟩offside⟨/ALT0⟩
⟨ALT1 CASE=”b”⟩offside, player, position⟨/ALT1⟩
⟨ALT2 CASE=”b”⟩offside, soccer⟨/ALT2⟩
⟨ALT3 CASE=”b”⟩soccer, offside, rule⟨/ALT3⟩
⟨USER⟩2nd year undergraduate student, male, 4 years of search experience, skill level 3, familiarity level 5⟨/USER⟩
⟨/TOPIC⟩

(b) An English translation of a sample topic

Figure 1. A sample topic for the Informational Retrieval Task and its English translation

3.2.2 Topic Creation Strategies

We applied the following strategies when creating the topics.

- All the topics were created without using any search systems or any relevance assessment.
- We instructed in advance not to create topics that depend strongly on time or change in time, although we understand that such topics are important in considering the user’s needs against the real Web. For instance, we discarded the topic, such as “I want to know the future match schedules of Hidetoshi Nakata —a Japanese famous soccer player—,” because the concept of ‘future’ depends strongly on time.
- The assessor described ⟨DESC⟩ in the topic under the following constraints: (1) The concepts or meanings of the terms specified in ⟨TITLE⟩ were included in ⟨DESC⟩, even though the terms themselves may not have appeared in ⟨DESC⟩; and (2) The ⟨DESC⟩ should have fundamentally included the scope that the topic indicated, avoiding a large gap between the scope of the ⟨DESC⟩ and that of the ⟨NARR⟩.

  These considerations were imposed because the systems often performed searches using the ⟨TITLE⟩ and/or ⟨DESC⟩, while the assessor judged the relevance on the basis of the scope of the ⟨NARR⟩.

267 topics were created by assessors. Then, we discarded inappropriate topics according to previously mentioned strategies. For the topics that were strongly similar to each other, we kept one of them and discarded the rest. We also discarded topics that were strongly similar to the ones that were created at the NTCIR-3 WEB. Consequently, we used the remaining 219 topics for the next step, ‘shallow pooling,’ which will be described in Section **3.3.1**.

3.3 Pooling

3.3.1 Topic Selection and Shallow Pooling

All the topics were created without using any search systems or any relevance assessment, as mentioned in Section **3.2.2**. Therefore, some of them were not suitable for use in a comparison of retrieval effectiveness. Therefore, we applied the following steps to discard inappropriate topics such as those with few relevant documents.

First, we investigated the search results of an our search system to discard inappropriate topics before delivering topics. As a result, 153 topics were selected for the formal run, and we delivered them to the participants.

Second, we performed ‘shallow pooling,’ which is a sampling method that takes the 10 highest-ranked documents from each run result submitted by a participant [5]. By assessing the relevance of each document included in the ‘shallow pool,’ we discarded 27 topics and used the remaining 126 topics for the next step, ‘deep pooling,’ which will be described in Section **3.3.2**.

3.3.2 Deep Pooling

Using the resulting topics of the shallow pooling, we perform ‘deep pooling,’ which took the potentially large number of highly ranked documents from each run result and merged them, as in the pooling methods previously used in conventional information retrieval evaluation workshops [10, 7]. Through the pooling stage, we obtain a subset of the document data, called the ‘pool,’ which was used to estimate the relevant documents included in the document data for the evaluation of the Informational Retrieval Task.

Using the result of shallow pooling, we divide the 128 topics into two groups: 53 and 75 topics. In the

pooling task using the 53 topics, we took the top 100 ranked documents from each run results (Pool P_1). Using the 75 topics, we took the only top 20 ranked documents from each run results (Pool P_2). We will evaluate the Informational Retrieval Task under the assumption of User-model U_1, as being described in Section **4**, using relevance assessment result for Pool P_1, and evaluate under the User-model U_2 using ones for both Pools P_1 and P_2.

We did not perform any additional manual searches to improve the comprehensiveness of relevant documents [7]; however, we attempted to improve the comprehensiveness of the pool by the following two ways: (i) adding run results using several baseline systems, and (ii) adding run results using various query expressions that were specified by several persons for the same information needs and that were extracted the first appeared term in the ⟨TITLE⟩ field of the topic.

3.3.3 Ranking of Pooled Documents

Using the result of Section **3.3.2**, we performed ranking of the pooled documents in the following manner:

(1) Iterate the following procedure, starting when $n = 1$ and stopping when $n \geq 100$ for Pool P_1 or $n \geq 20$ for Pool P_2: (i) Take the n-th ranked document from each run results, and randomly arrange them in a list; and then (ii) Take the $(n + 1)$-th ranked document from each run results, and randomly arrange them, following the list above.

(2) Manipulate the following procedure from the top-ranked document to the bottom-ranked in the list obtained from (1): If a document appear in a duplicated document group, move up the rest of documents in the group, following that document in the list.

The duplicated document groups were specified using completely duplicated pages, and using result of automatic detection of content duplication, which will be described in **3.3.4**. Procedure (2) was used for assessment of content duplication, which will be described in Section **3.4.4**, and motivated for the purpose of an evaluation considering non-redundancy, which will be described in Section **4.2**.

3.3.4 Automatic Detection of Content Duplication

After generating document pools each of which corresponds to the distinctive topic, possible duplications were detected using the following procedure.

First, exact duplications —except for their URL's— were removed from the pool so that none of the documents were identical to each other. The detected duplications were registered to a 'content duplication candidate list' that was later checked by human assessors in the manner as described in in Section **3.4** . Next, non-text documents were identified using UNIX 'file' command. Those documents were excluded from further consideration though they were still the subjects for relevance assessment.

Then, all the HTML tags, comments, and explicitly declared scripts were removed, and the documents were segmented into words using morphological analyzer 'ChaSen version 2.3.3' [2] . Here, EUC-converted documents included in the NW100G-01 dataset were used. Only the top 40 kilobytes were considered when the total length of the document exceeded the upper limit.

Finally, using suffix array-based clustering, all the document groups that satisfied the following conditions were enumerated and registered to a content duplication candidate list: (i) At least 100 consecutive words were shared by all the members of the group; (ii) The difference of the document lengths, in terms of the total number of words, was smaller than 0.5 of the largest document.

3.4 Relevance Assessment

Pooled documents that were composed of the highly ranked search results submitted by each participant were considered to be the relevant document candidates. Human assessors judged the relevance of each document in the pool using an assessment support system described in Section **3.4.1**, assuming the multiple document models described in Section **3.4.2**.

At that time, the assessors judged the 'multi-grade relevance' as highly relevant, fairly relevant, partially relevant or irrelevant, as described in Section **3.4.3**. Moreover, the assessors also made an assessment of content duplication, as described in Section **3.4.4**.

3.4.1 Assessment Support System

The assessment support system that we used at the NTCIR-4 WEB ran on our HTTP server, and was available through CGIs. That was basically the same as the one used at the NTCIR-3 WEB, but the usability had been improved. All the pooled documents to be assessed were ranked in the manner described in **3.3.3**, and converted to almost plain text. Individual documents to be judged and their out-linked pages that were included in the pool were listed. When assessors judged the relevance of a document, they basically browsed its converted text and that of the out-linked pages; however, they could refer to the non-converted pages that had the same contents.

3.4.2 Document Models

Web pages are represented in various ways, so that in one example, an 'information unit' on the Web could

[2] ⟨http://chasen.aist-nara.ac.jp⟩

be hyper-linked pages, while in another, it could be an individual page, or a passage included on a page.

Assuming an information unit to be a page, a 'hub page' [8] that gives out-links to multiple 'authority pages' must be judged as irrelevant if these do not include sufficient relevant information in them. However, in the Web environment, this type of hub page is sometimes more useful for the user than the relevant pages defined by the assumption.

Therefore, We assumed the following two document models:

One-click-distance document model This was where the assessor judged the relevance of a page when he/she could browse the page and its 'out-linked pages' that satisfied some of the conditions, but not all of the out-linked pages. The out-linked pages indicate pages that are connected from a certain page whose anchor tags describe the URLs of the out-linked pages.

We imposed the following conditions on the out-linked pages to be browsed: that the out-linked pages should be included in the pool, assuming that most of the relevant documents may be included in the pool[3].

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

3.4.3 Multi-Grade Relevance

The assessors judged the 'Multi-Grade Relevance' of the individual pooled documents as: highly relevant, fairly relevant, partially relevant or irrelevant. Here, the number of documents corresponding to each grade were not controlled —for example, the assessor did not care if the number of highly relevant documents were very small—. In this paper, we denote the highly relevant, fairly relevant, and partially relevant documents as being a 'relevant document' as long as we do not have to specify the grade of relevance.

The assessors judged the relevance of the pooled documents only on the basis of the information given in Japanese or English. The documents included in the document data seemed to be described in various languages, because we had not discarded documents with page data described in languages other than Japanese or English from the document data. If a part of the pooled documents were entirely described in languages other than Japanese or English, the assessors must have judge this kind of documents as being irrelevant.

[3]Pool P_2 described in Section **3.3.2** was collected using top 20 ranked documents from each run results; however, we separately collected using top 100 ranked documents from each run results to specify the out-linked pages to be browsed even in this case.

3.4.4 Assessment of Content Duplication

The assessors judged the content duplication of relevant documents using the result of the automatic detection of content duplication that was described in Section **3.3.4**.

While they assessed the content duplication, they judged from the viewpoint of information needs stated in each topic description. Even if most of the contents of documents were similar to each other, these documents might be deemed as not being duplicated when the difference of them were strongly related to the information needs. When the contents of a document were judged as a part of the contents of another one and as being related to the information needs, these two documents might be deemed as being duplicated.

The assessment result will be used for an evaluation considering non-redundancy, as being described in Section **4.2**.

4 Evaluation Measures

4.1 User Models and Evaluation Measures

We supposed two kinds of user models for evaluations: (i) the user model U_1 where a user attempted to comprehensively find documents relevant to his/her information needs, such as in the Survey Retrieval Tasks at the NTCIR-3 WEB, and (ii) the user model U_2 where the user requires just one or only a few relevant documents, so the precision of the highly ranked search results is emphasized, such as in the Target Retrieval Task at the NTCIR-3 WEB [5, 4]. After run result submission, we divided the topics into two groups for evaluation of (i) and (ii), respectively, as mentioned in Section **3.3.2**.

In evaluating the run results of each participant's search engine system, we focused on up to 1,000 top-ranked documents for Model U_1, and up to 20 top-ranked documents for Model U_2.

For Model U_1, we applied the two types of evaluation measures: (i) those based on precision and/or recall, and (ii) those with discounted cumulative gain ('DCG') [6]. For Model U_2, we applied the three types of measures: the aforementioned measures in (i) and (ii), and weighted reciprocal rank measure ('WRR') ((iii)) [5, 4].

Although the one-click-distance document model was partly applied in the relevance assessment, as described in Section **3.4.2**, 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 was 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 doc-

ument sets, according to which of the two document models was used.

4.2 An Evaluation Method Considering Non-redundancy

When duplicate pages or closely linked pages appear in the Web search engine results, they are often unwelcome for users. We carried out an evaluation considering non-redundancy of pages [5, 4], as described below, using the related document groups as the result of the assessment of content duplication that were described in Section **3.4.4**.

- For the document, comprising the related document group, that first appeared in each run result list, we treated this kind of document as it is.
- For the other related documents, we treated them as irrelevant (or partially relevant) although they were judged as relevant.

Consequently, run results that contained the duplicated documents or closely linked ones were expected to pay a penalty.

We designed this evaluation method by supposing it to be combined with the precision-recall-related measures or the DCG measure, which were described in Section **4.1**. The evaluation results will be described in another article.

4.3 An Evaluation Method Considering Diversity of Query Expressions

Each topic statement includes (i) various query expressions specified by several persons for the same information needs as ⟨ALT1⟩, ⟨ALT2⟩ and ⟨ALT3⟩, as well as ⟨TITLE⟩, and (ii) a query expressed by a single term having a vague or broader meaning as ⟨ALT0⟩, by extracting the first appeared term in the ⟨TITLE⟩ field of the topic, as described in Section **3.2.1**.

We mentioned in Section **3.3.2** that we expected these queries to improve the comprehensiveness of the pool. We also expect that those queries can be used for some kinds of evaluation methods, such as an evaluation method from a viewpoint of robustness of retrieval performance. Details will be forthcoming in another article.

4.4 An Evaluation Method based on Users' Sense

Ohtsuka *et al.* proposed a user-oriented criterion for evaluating Web search systems, considering users' search behavior. As organizers of the NTCIR-4 WEB, they attempted to evaluate the Informational Retrieval Task using a part of the data of submitted run results and the topics, comparing the proposed criterion with conventional evaluation methods by measuring the time spent on search as the users' satisfaction degree. The details can be found in Reference [9] included in this volume.

5 Evaluation Results

5.1 Summary of Participation

Five groups, listed below in alphabetical order of affiliations, submitted their completed run results.

- Hokkaido University
- National Institute of Informatics, the University of Tokyo, and KYA group
- Osaka Kyoiku University
- Toyohashi University of Technology
- University of Tsukuba, Nagoya University, and National Institute of Advanced Industrial Science and Technology

We asked three research groups to submit run results along with those of the participants in an attempt to improve the comprehensiveness of the pool and as baseline data. The first group ('GRACE') of them were not participating in the NTCIR-4 WEB, however, we asked them because they performed excellently at the NTCIR-3 WEB. The second group ('K3100') were participating in only the Navigational Retrieval Task at the NTCIR-4 WEB. They also participated in NTCIR-3 WEB. The third group ('TKB') were participating in the Informational Retrieval Task so that they submitted additional run results. The 'GRACE' and 'K3100' groups performed their search systems used at the NTCIR-3 WEB, not making special efforts, such as parameter tuning, to cope with the Informational Retrieval Task at the NTCIR-4 WEB. The organizers also submitted run results using two types of search systems ('ORGREF' and 'NAMAZU' as described below) to improve the comprehensiveness of the pool and as baseline data.

The individual participating groups pursued various objectives. We summarize them as follows:

(1) Participation

DBLAB Experimented with **(i)** the text retrieval system based on the probabilistic model that is similar to 'OKAPI' model using both word-based and phrase-based indexes, and **(ii)** the automatic clarification of Boolean queries.

OKSAT Experimented with character n-gram indexing and retrieval system based on a probabilistic model. Additional experiments were performed on the basis of link structure analysis, such as for re-ranking search results using 'dynamic PageRank,' although the run results submission was missed on our evaluation.

Table 1. Summary of run result submission

Participation						Submission at the request of organizers or submission by organizers					
RunID	QueryMethod	TopicPart	Cont	Link	Anchor	RunID	QueryMethod	TopicPart	Cont	Link	Anchor
DBLAB-tt-01	automatic	TITLE w/ C	yes	no	no	GRACE-tt-01	automatic	TITLE w/o C	yes	no	no
DBLAB-tt-02	automatic	TITLE w/o C	yes	no	no	GRACE-tt-02	automatic	TITLE w/o C	yes	no	no
DBLAB-ds-01	automatic	DESC	yes	no	no	GRACE-ds-01	automatic	DESC	yes	no	no
DBLAB-ds-02	automatic	DESC	yes	no	no	GRACE-ds-02	automatic	DESC	yes	no	no
OKSAT-tt-01	automatic	TITLE w/o C	yes	no	no	K3100-tt-01	automatic	TITLE w/o C	yes	no	yes
OKSAT-it-02	interactive	TITLE, ALT1-3 w/ C	yes	no	no	K3100-tt-02	automatic	TITLE w/o C	no	no	yes
OKSAT-it-03	interactive	TITLE, ALT1-3 w/ C	yes	no	no	K3100-ds-01	automatic	DESC	yes	no	yes
OKSAT-ds-01	automatic	DESC	yes	no	no	K3100-ds-02	automatic	DESC	no	no	yes
R2D2-tt-01	automatic	TITLE w/o C	yes	no	no	NAMAZU-tt-01	automatic	TITLE w/o C	yes	no	no
R2D2-ds-01	automatic	DESC	yes	no	no	NAMAZU-tt-02	automatic	TITLE w/ C	yes	no	no
SSTUT-tt-01	automatic	TITLE w/o C	no	no	yes	NAMAZU-ds-01	automatic	DESC	yes	no	no
SSTUT-tt-02	automatic	TITLE w/o C	yes	no	no	ORGREF-tt-01	automatic	TITLE w/ C	yes	no	no
SSTUT-tt-03	automatic	TITLE w/o C	yes	no	yes	ORGREF-tt-02	automatic	TITLE w/ C	yes	no	no
SSTUT-ds-01	automatic	DESC	no	no	yes	ORGREF-tt-03	automatic	TITLE w/ C	yes	no	no
SSTUT-ds-02	automatic	DESC	yes	no	no	ORGREF-tt-04	automatic	TITLE w/ C	yes	no	no
SSTUT-ds-03	automatic	DESC	yes	no	yes	ORGREF-tt-05	automatic	TITLE w/ C	yes	no	no
TKB-tt-01	automatic	TITLE w/o C	yes	no	no	ORGREF-tt-06	automatic	TITLE w/ C	yes	no	no
TKB-tt-02	automatic	TITLE w/o C	yes	no	no						
TKB-tt-03	automatic	TITLE w/o C	yes	yes	no						
TKB-tt-04	automatic	TITLE w/o C	yes	yes	no						
TKB-ds-01	automatic	DESC	yes	no	no						
TKB-ds-02	automatic	DESC	yes	no	no						
TKB-ds-03	automatic	DESC	yes	yes	no						
TKB-ds-04	automatic	DESC	yes	yes	no						
TKB-it-01	interactive	DESC, BACK, TERM	yes	no	no						
TKB-it-02	interactive	DESC, BACK, TERM	yes	no	no						

RunID: Indicates the identification codes of the system run results. Each one starts with the group ID.
QueryMethod: Indicates 'automatic' or 'interactive'. 'Automatic' indicates a run without any human intervention during query processing and search; 'interactive' indicates a run other than 'automatic'.
TopicPart: Indicates the part of the topic used. 'w/ C' and 'w/o C' indicate if the system used Boolean operators that were specified as 'CASE' attribute or not, respectively.
Cont: Indicates whether or not the system used textual contents of web documents for indexing.
Link: Indicates whether or not the system used link information in Web documents. The notation 'yes' indicates that the links and contents were used; 'no' indicates that only contents were used.
Anchor: Indicates whether or not the system used anchor text for indexing.

R2D2 Experimented with 'Relevance-based Superimposition (RS) model,' on the basis of vector space model, which modified document feature vectors using document clusters. The document clusters were gathered using automatically extracted keywords.

SSTUT Experimented with **(i)** score merging based on anchor text that pointed to a page and textual contents of the page. The retrieval method was based on 'OKAPI' model. Additional experiments were performed, such as **(ii)** query term reweighting considering the entropy-based weights on the 'virtual document space,' and **(iii)** reranking of search results using 'literal matching aided link analysis'; however, the run results submission of (ii) and (iii) were missed on our evaluation.

TKB Experimented with **(i)** the text retrieval system based on 'OKAPI' model and partially using pseudo-relevance feedback, **(ii)** a speech-driven retrieval system where speech recognition and the text retrieval modules, as described in (i), were integrated, and **(iii)** the text retrieval system as described in (i) and using 'PagaRank' re-ranking. System (iii) was used for additional experiments, where the run results were submitted at the request of organizers.

(2) Submission at the request of organizers

GRACE Used the system implemented at NTCIR-3 WEB using the text retrieval method based on probabilistic model that is similar to 'OKAPI' model, and partially using pseudo-relevance feedback.

K3100 Used the system implemented at NTCIR-3 WEB using a retrieval method based on anchor text that pointed to a web page or its web site as well as textual contents of the web page.

(3) Submission by organizers

NAMAZU Performed by the organizers using a freely available tool[4], where searching with/without the Boolean operators and ranking by tf-idf were available.

ORGREF Performed by the organizers using a Boolean-type search system, where searching by the presence of proximity and ranking by tf-idf were available.

Summaries of the run result submissions of each participating group can be found in **Table 1**, and most of the details can be found in papers of the participating groups in this volume.

[4]We used 'Namazu' as a search system, which is freely available at ⟨http://www.namazu.org/⟩.

5.2 Experimental Conditions

In evaluating the run results, we used combinations of $\{UM_1, UM_2\} \times \{DM_1, DM_2\} \times \{RL_1, RL_2\}$, which were defined as follows:

User Models (as described in Section **4.1**)

(UM_1) 'Survey'-type Is assuming the model where the user attempted to comprehensively find documents relevant to his/her information needs.

(UM_2) 'Target'-type Is assuming the model where the user requires just one or only a few relevant documents at the highly ranked documents.

Document Models (as described in Section **3.4.2**)

(DM_1) One-click-distance document model

(DM_2) Page-unit document model

Relevance Levels

(RL_1) Rigid relevance level For the precision-recall-related measures at the Rigid relevance level, we considered the document to be relevant if it was highly relevant or fairly relevant, and otherwise considered it to be irrelevant.

For the DCG evaluation measures as mentioned in Section **4**, we set magnitude of the gain as $(g_S, g_A, g_B) = (3, 2, 0)$, where g_S, g_A and g_B indicate the magnitude of the gain for highly relevant, fairly relevant or partially relevant documents, respectively, in the equation of DCG calculation [6].

We simply computed WRR evaluation measures as mentioned in Section **4** using $(\delta_h, \delta_a, \delta_b) = (1, 1, 0)$, and $(\beta_h, \beta_a, \beta_b) = (\infty, \infty, \infty)$.

(RL_2) Relaxed relevance level For the precision-recall-related measures at the Relaxed relevance level, we considered the document to be relevant if it was highly relevant, fairly relevant or partially relevant, and otherwise considered it to be irrelevant.

For the DCG evaluation measures, we set magnitude of the gain as $(g_S, g_A, g_B) = (3, 2, 1)$.

We simply computed WRR evaluation measures using $(\delta_h, \delta_a, \delta_b) = (1, 1, 1)$, and $(\beta_h, \beta_a, \beta_b) = (\infty, \infty, \infty)$.

5.3 Summary of Evaluation Results

In this paper, we used 35 topics using document pools from highly ranked 100 documents —assuming the Survey-type, UM_1—, and 45 topics using document pools from top 20 documents —assuming the Target-type, UM_2—. These 35 topics were also used for the evaluation under UM_2; therefore 80 topics[5]could be used for the evaluation under UM_2.

We computed the effectiveness of individual run results as shown in Section **5.1** using the respective evaluation measures described in Section **4** and using the conditions as described in Section **5.2**. Selected evaluation results of the Survey Retrieval Tasks and the Target Retrieval Task are shown in **Tables 2** and **3**, respectively. In each task and part of the topic used, the run ID codes denoted in the tables are ranked in order of the average precision in the One-click-distance document model, DM_1 and the Rigid relevance level, RL_1 for the Survey-type evaluation; and in order of the precision at 10 document-level in DM_1 and RL_1 for the Target-type evaluation. In the tables, each evaluation values were averaged over all the 35 topics for the Survey-type evaluation, or were averaged over all the 80 topics for the Target-type evaluation. We omit evaluation results under the Page-unit document model in this paper.

Selected recall-precision and DCG curves are also shown in **Figures 3** and **4**. In these graphs, all the run results were performed 'automatically'. Some 'Interactive' run results were submitted, but there were too few of them. The terminologies of 'automatic' and 'interactive' are explained in Section **2**. In each graph, the explanatory notes report the run ID codes, which are ranked in order of the average precision. In the graphs, each of the run ID codes identifies the best run selected for the individual participating group.

6 Conclusions

We have described an overview of the Informational Retrieval Task 2 of the WEB Task at the Fourth NTCIR Workshop. To evaluate the task, we used the 100-gigabyte document dataset that was used at the previous workshop experiments and were mainly gathered from the '.jp' domain. The topics were designed to resemble real Web retrieval tasks. Relevance judgments were performed on the retrieved documents written in Japanese or English, in part, by considering the effects of linked pages. The system results submitted by the participants were evaluated according to various measures.

We also investigated the evaluation methods considering non-redundancy of contents and diversity of queries, however, the the evaluation using them is currently in progress and so will be described the results of them in another article.

[5] As we mentioned in Section **3.3.2**, the number of the available topics should be at most 128, however, we have not completed the evaluation using all of them. Therefore, we describe the evaluation results using only the 80 topics in this paper.

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 Informational Retrieval Task 2 of the WEB Task at the Fourth NTCIR Workshop. We would like to greatfully acknowledge Dr. Yasushi Ogawa and Dr. Hideo Itoh at RICOH Co. Ltd., Dr. Toshikazu Fukushima, Mr. Kenji Tateishi and Mr. Hideki Kawai at NEC Corporation, and Dr. Atsushi Fujii at University of Tsukuba for providing their system results. We also would like thank Professor Jun Adachi at National Institute of Informatics, and Dr. Hayato Yamana at Waseda University for their useful advice.

References

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

[2] N. Craswell and D. Hawking. Overview of the TREC-2002 Web Track. In *Proc. of 11th Text REtrieval Conference (TREC 2002)*, Gaithersburg, USA, 2002.

[3] N. Craswell and D. Hawking. Overview of the TREC 2003 Web Track. In *Proc. of 12th Text REtrieval Conference (TREC 2003)*, Gaithersburg, USA, 2003.

[4] 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*, E86-D(9):1804–1813, 2003.

[5] K. Eguchi, K. Oyama, E. Ishida, N. Kando, and K. Kuriyama. Overview of the Web Retrieval Task at the Third NTCIR Workshop. Technical Report NII-2003-002E, National Institute of Informatics, 2003.

[6] K. Järvelin and J. Kekäläinen. IR evaluation methods for retrieving highly relevant documents. In *Proc. of 23rd Annual International ACM SIGIR Conference*, pages 41–48, Athens, Greece, 2000.

[7] N. Kando, K. Kuriyama, T. Nozue, K. Eguchi, H. Kato, and S. Hidaka. Overview of IR tasks at the First NTCIR Workshop. In *Proc. of 1st NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, pages 11–22, Tokyo, Japan, 1999.

[8] J. Kleinberg. Authoritative sources in a hyperlinked environment. In *Proc. of 9th ACM SIAM Symposium on Discrete Algorithms*, San Francisco, USA, 1998.

[9] T. Ohtsuka, K. Eguchi, and H. Yamana. An evaluation method of web search engines based on users' sense. In *Proceedings of the Fourth NTCIR Workshop on Research in Information Access Technologies: Information Retrieval, Question Answering and Summarization*, Tokyo, Japan, 2004.

[10] E. Voorhees and D. K. Harman. Overview of the Sixth Text REtrieval Conference (TREC-6). In *Proc. of 6th Text REtrieval Conference (TREC-6)*, pages 1–24, Gaithersburg, USA, 1997.

Appendix: System Description Form

Each participating group was expected to submit a concise description of each run according to the following format:

- ⟨Subtask⟩ is fixed to 'A' in the Informational Retrieval Task.
- ⟨RunID⟩ identifies each run result in the manner of '⟨groupid⟩-⟨tagid⟩-⟨serialnumber⟩.res,' *e.g.*, 'orgref-ds-01.res,' where the ⟨groupid⟩ indicates the group identification. The ⟨tagid⟩ indicates the part of the topic used for searching, such as 'ds' (description), 'tt' (title), 'a0,' $\cdots$,'a3' (alternative queries). The ⟨serialnumber⟩ indicates the serial number of the run.
- ⟨QueryMethod⟩ indicates whether the run is 'automatic' or 'interactive'. The 'automatic' and the 'interactive' indicate runs without any human intervention during search process, and all runs other than 'automatic,' respectively
- ⟨TopicPart⟩ specifies the part of the topic statement used for searching. This tag also specifies whether each participating group used 'CASE' or not, *e.g.*, 'DESC,' 'TITLE w/ CASE,' 'TITLE w/o CASE,' 'ALT0,' 'ALT1 w/ CASE,' 'ALT1 w/o CASE,' etc.
- ⟨QueryUnit⟩ specifies the unit of the query used for searching, *e.g.*, character, word, phrase, etc.
- ⟨QueryExpan⟩ specifies the techniques used to expand query, *e.g.*, pseudo-relevance feedback, no query expansion, etc.
- ⟨LinkInfo⟩ specifies whether or not link information was used for indexing or searching, *e.g.*, link information only, link and contents information, contents only, etc.
- ⟨Anchor⟩ specifies whether or not anchor text was used for indexing or searching, *e.g.*, used for indexing the document that includes the anchor text, used for indexing the out-linked document, not used, etc.
- ⟨IRModel⟩ specifies the information retrieval model, *e.g.*, Boolean model, vector space model, probabilistic model, etc.
- ⟨Ranking⟩ specifies the technical factor for computing ranking scores, *e.g.*, tf, tf-idf, mutual information, document length, PageRank, etc.
- ⟨IndexUnit⟩ specifies the unit of index, *e.g.*, character, bi-character, word, bi-word, phrase, the name of the HTML tags used, link structure, etc.
- ⟨IndexTech⟩ specifies the techniques used to process index terms, *e.g.*, morphology, stemming, POS, etc.
- ⟨IndexStruc⟩ specifies the index structure, *e.g.*, PAT, inverted file, signature file, etc.

⟨SYSDESC⟩
⟨SUBTASK⟩Subtask⟨/SUBTASK⟩
⟨RUNID⟩RunID⟨/RUNID⟩
⟨TOPICPART⟩TopicPart⟨/TOPICPART⟩
⟨QUERYMETHOD⟩QueryMethod⟨/QUERYMETHOD⟩
⟨QUERYUNIT⟩QueryUnit⟨/QUERYUNIT⟩
⟨QUERYEXPAN⟩QueryExpan⟨/QUERYEXPAN⟩
⟨LINKINFO⟩LinkInfo⟨/LINKINFO⟩
⟨ANCHOR⟩Anchor⟨/ANCHOR⟩
⟨IRMODEL⟩IRModel⟨/IRMODEL⟩
⟨RANKING⟩Ranking⟨/RANKING⟩
⟨INDEXUNIT⟩IndexUnit⟨/INDEXUNIT⟩
⟨INIDEXTECH⟩IndexTech⟨/INIDEXTECH⟩
⟨INDEXSTRUC⟩IndexStruc⟨/INDEXSTRUC⟩
⟨DUPREDUCT⟩DupReduct⟨/DUPREDUCT⟩
⟨FILTERING⟩Filtering⟨/FILTERING⟩
⟨RESOURCE⟩Resource⟨/RESOURCE⟩
⟨PRIORITY⟩Priority⟨/PRIORITY⟩
⟨RUNTIME⟩RunTime⟨/RUNTIME⟩
⟨INDEXTIME⟩IndexTime⟨/INDEXTIME⟩
⟨NOTE⟩Note⟨/NOTE⟩
⟨/SYSDESC⟩

Figure 2. Format of System Description

- ⟨DupReduct⟩ specifies the techniques used to reduce content duplication.
- ⟨Filtering⟩ specifies the filtering method used for distilling useful pages other than junk pages, *e.g.*, Kleinburg and Chakrabarti's topic distillation, Web pages selection using internet directory, SPAM filtering using SPAM-like words or patterns, etc.
- ⟨Resource⟩ specifies the the external resources used for indexing, filtering, or searching, other than the data provided by organizers, *e.g.*, an internet directory, a training dataset, etc.
- ⟨Priority⟩ specifies the priority rank to each of four RunIDs, *e.g.*, 1, 2, or 'RunID:1, RunID:2 when designating more than one runids at once.
- ⟨RunTime⟩ optionally specifies the averaged seconds consumed for searching.
- ⟨IndexTime⟩ optionally specifies the averaged hours consumed for indexing
- ⟨Note⟩ optionally specifies any other additional information.

Each system description should be flanked by '⟨SYSDESC⟩' and '⟨/SYSDESC⟩,' as shown in **2**. Each participating group was encouraged to describe all the items in detail and concretely, not limited to the examples indicated above.

Table 2. Selected results of the survey-type evaluation

Query-Method	TopicPart	RunID	Cont	Link	Anchor	DM_1&RL_1				DM_1&RL_2			
						a-prec	r-prec	dcg(100)	dcg(1K)	a-prec	r-prec	dcg(100)	dcg(1K)
automatic	TITLE w/ C	DBLAB-tt-01	yes	no	no	0.2189	0.2455	13.0961	23.7245	0.2438	0.2881	14.8590	27.5607
automatic	TITLE w/o C	DBLAB-tt-02	yes	no	no	0.2155	0.2421	12.8458	23.4989	0.2401	0.2814	14.5849	27.3049
automatic	TITLE w/o C	GRACE-tt-02	yes	no	no	0.1985	0.2376	12.7676	25.0382	0.2164	0.2787	14.4183	28.7361
automatic	TITLE w/o C	GRACE-tt-01	yes	no	no	0.1716	0.2062	12.0872	23.3912	0.1930	0.2491	13.7638	27.1963
automatic	TITLE w/o C	SSTUT-tt-02	yes	no	no	0.1439	0.1813	10.7419	20.7585	0.1672	0.2199	12.3679	24.3188
automatic	TITLE w/o C	R2D2-tt-01	yes	no	no	0.1417	0.1698	10.2175	22.3168	0.1602	0.2150	11.6368	25.5065
automatic	TITLE w/ C	ORGREF-tt-06	yes	no	no	0.1328	0.1766	10.1938	20.8703	0.1336	0.1828	11.0203	23.3080
automatic	TITLE w/o C	TKB-tt-01	yes	no	no	0.1255	0.1671	8.7843	19.1973	0.1635	0.2083	10.4688	22.7578
automatic	TITLE w/o C	K3100-tt-01	yes	no	yes	0.1237	0.1673	9.9174	19.5139	0.1321	0.1917	11.2491	22.1067
automatic	TITLE w/ C	ORGREF-tt-05	yes	no	no	0.1160	0.1574	9.2942	18.9844	0.1270	0.1809	10.6435	21.6688
automatic	TITLE w/o C	TKB-tt-02	yes	no	no	0.1154	0.1522	9.2896	19.6625	0.1397	0.1966	10.8349	22.8807
automatic	TITLE w/o C	OKSAT-tt-01	yes	no	no	0.1075	0.1502	9.6178	17.4799	0.1007	0.1542	10.3583	18.7156
automatic	TITLE w/o C	SSTUT-tt-03	yes	no	yes	0.0973	0.1476	8.5827	16.5107	0.1124	0.1641	10.1501	19.6845
automatic	TITLE w/ C	ORGREF-tt-01	yes	no	no	0.0967	0.1368	8.4414	18.0627	0.0965	0.1404	9.0616	19.5141
automatic	TITLE w/ C	ORGREF-tt-02	yes	no	no	0.0938	0.1311	8.2950	17.8701	0.0942	0.1423	8.9536	19.2982
automatic	TITLE w/ C	ORGREF-tt-03	yes	no	no	0.0860	0.1258	8.5916	15.0109	0.0894	0.1445	9.6444	17.0943
automatic	TITLE w/ C	ORGREF-tt-04	yes	no	no	0.0860	0.1247	8.4063	14.7732	0.0905	0.1462	9.4716	16.8609
automatic	TITLE w/o C	TKB-tt-03	yes	yes	no	0.0838	0.1194	6.9355	16.0628	0.1120	0.1556	8.2200	19.0702
automatic	TITLE w/o C	TKB-tt-04	yes	yes	no	0.0521	0.0844	5.7428	14.5163	0.0642	0.1094	6.5684	16.6758
automatic	TITLE w/ C	NAMAZU-tt-02	yes	no	no	0.0275	0.0497	3.7798	4.6225	0.0351	0.0620	4.3648	5.5060
automatic	TITLE w/o C	SSTUT-tt-01	no	no	yes	0.0252	0.0520	3.5812	6.5065	0.0263	0.0649	4.3630	8.0604
automatic	TITLE w/o C	NAMAZU-tt-01	yes	no	no	0.0182	0.0359	2.3576	6.9389	0.0229	0.0490	2.8013	8.3505
automatic	TITLE w/o C	K3100-tt-02	no	no	yes	0.0014	0.0032	0.3161	0.3459	0.0013	0.0037	0.5323	0.5659
automatic	DESC	GRACE-ds-02	yes	no	no	0.1948	0.2426	12.8608	24.0043	0.2158	0.2784	14.5695	27.9739
automatic	DESC	DBLAB-ds-01	yes	no	no	0.1895	0.2189	11.7476	21.1413	0.2115	0.2541	13.2986	24.7766
automatic	DESC	DBLAB-ds-02	yes	no	no	0.1893	0.2184	11.7139	21.1482	0.2127	0.2583	13.2510	24.8346
automatic	DESC	GRACE-ds-01	yes	no	no	0.1565	0.2054	11.6463	21.6387	0.1778	0.2429	13.2193	25.3052
automatic	DESC	SSTUT-ds-02	yes	no	no	0.1129	0.1623	8.7351	17.4139	0.1319	0.1890	10.0772	20.4499
automatic	DESC	TKB-ds-01	yes	no	no	0.1011	0.1358	7.8213	17.4107	0.1351	0.1932	9.5019	21.0610
automatic	DESC	R2D2-ds-01	yes	no	no	0.0926	0.1215	7.8792	16.6951	0.1074	0.1557	8.9381	19.4037
automatic	DESC	SSTUT-ds-03	yes	no	yes	0.0865	0.1289	7.9024	15.4373	0.1007	0.1482	9.1636	18.2146
automatic	DESC	TKB-ds-02	yes	no	no	0.0785	0.1144	7.3461	15.8315	0.1017	0.1621	8.7296	18.9528
automatic	DESC	K3100-ds-01	yes	no	yes	0.0730	0.1052	6.7854	13.8143	0.0843	0.1227	7.8836	16.1267
automatic	DESC	TKB-ds-03	yes	yes	no	0.0577	0.0941	5.4075	13.4835	0.0823	0.1250	6.7012	16.5106
automatic	DESC	OKSAT-ds-01	yes	no	no	0.0400	0.0675	4.5103	9.1623	0.0511	0.0871	5.2731	11.0081
automatic	DESC	TKB-ds-04	yes	yes	no	0.0381	0.0638	4.1513	10.9281	0.0472	0.0874	4.8448	12.9612
automatic	DESC	SSTUT-ds-01	no	no	yes	0.0195	0.0470	3.1736	5.4825	0.0224	0.0607	3.8146	6.7941
automatic	DESC	NAMAZU-ds-01	yes	no	no	0.0080	0.0194	1.2219	3.5551	0.0109	0.0305	1.4985	4.5468
automatic	DESC	K3100-ds-02	no	no	yes	0.0003	0.0018	0.2552	0.3049	0.0008	0.0030	0.4155	0.4729
interactive	TITLE, ALT1-3 w/ C	OKSAT-it-02	yes	no	no	0.1262	0.1568	10.0697	19.3559	0.1324	0.1715	11.0303	21.7653
interactive	DESC, BACK, TERM	TKB-it-01	yes	no	no	0.1037	0.1385	8.4546	18.3983	0.1210	0.1679	9.6035	20.9425
interactive	TITLE, ALT1-3 w/ C	OKSAT-it-03	yes	no	no	0.0978	0.1389	8.3147	18.2423	0.1060	0.1530	9.2317	20.5100
interactive	DESC, BACK, TERM	TKB-it-02	yes	no	no	0.0802	0.1225	7.6970	16.0945	0.0918	0.1443	8.7062	18.4387
mean						0.1006	0.1341	7.9437	15.9349	0.1141	0.1596	9.0718	18.4603

(In each query method and part of the topic used, the run ID codes are ranked in order of the average precision in DM_1 and RL_1.)

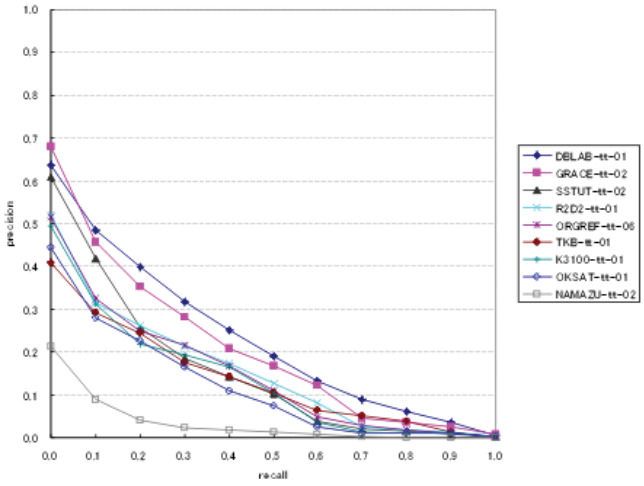
QueryMethod: Indicates 'automatic' or 'interactive'. 'Automatic' indicates a run without any human intervention during query processing and search; 'interactive' indicates a run other than 'automatic'.
TopicPart: Indicates the part of the topic used. 'w/ C' and 'w/o C' indicate if the system used Boolean operators that were specified as 'CASE' attribute or not, respectively.
RunID indicates the identification codes of the system run results, as shown in **Table 1**.
Cont: Indicates whether or not the system used textual contents of web documents for indexing.
Link: Indicates whether or not the system used link information in Web documents. The notation 'yes' indicates that the links and contents were used; 'no' indicates that only contents were used.
Anchor: Indicates whether or not the system used anchor text for indexing.
a-prec indicates the average precision (non-interpolated).
r-prec indicates the R-precision.
dcg(100) indicates the DCG value at the 100-document level.
dcg(1K) indicates the DCG value at the 1,000-document level.

Table 3. Selected results of the target-type evaluation

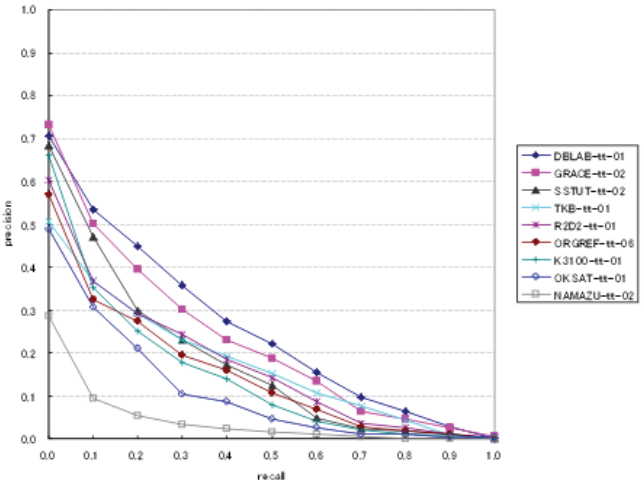
Query-Method	TopicPart	RunID	Cont	Link	Anchor	DM_1&RL_1				DM_1&RL_2			
						prec(10)	dcg(10)	wrr(10)	%nf(10)	prec(10)	dcg(10)	wrr(10)	%nf(10)
automatic	TITLE w/o C	GRACE-tt-02	yes	no	no	0.4175	5.2983	0.5975	0.1375	0.5038	5.7306	0.6586	0.1000
automatic	TITLE w/o C	GRACE-tt-01	yes	no	no	0.3888	5.0385	0.6316	0.1375	0.4825	5.5288	0.7401	0.0625
automatic	TITLE w/o C	DBLAB-tt-02	yes	no	no	0.3750	4.7569	0.5712	0.2625	0.4675	5.2313	0.6672	0.1875
automatic	TITLE w/o C	SSTUT-tt-02	yes	no	no	0.3250	4.1890	0.5207	0.2500	0.4188	4.6798	0.5916	0.1875
automatic	TITLE w/o C	R2D2-tt-01	yes	no	no	0.3163	3.9861	0.4928	0.2250	0.3938	4.4112	0.5932	0.1250
automatic	TITLE w/o C	SSTUT-tt-03	yes	no	yes	0.2975	3.9379	0.4945	0.2000	0.3850	4.4284	0.5913	0.1500
automatic	TITLE w/o C	TKB-tt-01	yes	no	no	0.2888	3.3980	0.3807	0.3875	0.3738	3.8465	0.4772	0.2875
automatic	TITLE w/o C	K3100-tt-01	yes	no	yes	0.2738	3.3620	0.4499	0.2375	0.3663	3.8697	0.5778	0.1250
automatic	TITLE w/o C	OKSAT-tt-01	yes	no	no	0.2675	3.3531	0.3930	0.3375	0.3313	3.6628	0.4633	0.2625
automatic	TITLE w/o C	TKB-tt-02	yes	no	no	0.2600	3.1257	0.4148	0.3125	0.3500	3.5799	0.5161	0.1875
automatic	TITLE w/o C	TKB-tt-03	yes	yes	no	0.1938	2.2374	0.2785	0.4750	0.2513	2.5467	0.3668	0.3750
automatic	TITLE w/o C	SSTUT-tt-01	no	no	yes	0.1488	1.9790	0.3126	0.5125	0.2075	2.3045	0.3908	0.4000
automatic	TITLE w/o C	TKB-tt-04	yes	yes	no	0.1338	1.6170	0.2294	0.5625	0.1825	1.8739	0.3166	0.4125
automatic	TITLE w/o C	NAMAZU-tt-01	yes	no	no	0.0775	0.9160	0.1345	0.7375	0.0975	1.0589	0.1957	0.6500
automatic	TITLE w/o C	K3100-tt-02	no	no	yes	0.0275	0.3711	0.0802	0.6250	0.0438	0.4705	0.1188	0.5625
automatic	TITLE w/ C	DBLAB-tt-01	yes	no	no	0.3725	4.7376	0.5712	0.2625	0.4650	5.2126	0.6677	0.1875
automatic	TITLE w/ C	ORGREF-tt-03	yes	no	no	0.2700	3.2517	0.4442	0.3000	0.3375	3.6527	0.5557	0.1750
automatic	TITLE w/ C	ORGREF-tt-04	yes	no	no	0.2700	3.2336	0.4260	0.3125	0.3525	3.7135	0.5442	0.1750
automatic	TITLE w/ C	ORGREF-tt-06	yes	no	no	0.2575	3.0610	0.3922	0.3125	0.3213	3.4123	0.4832	0.2000
automatic	TITLE w/ C	ORGREF-tt-05	yes	no	no	0.2400	2.9500	0.3613	0.3125	0.3188	3.4035	0.4947	0.2250
automatic	TITLE w/ C	ORGREF-tt-01	yes	no	no	0.1863	2.0663	0.2378	0.5125	0.2338	2.3129	0.3137	0.4125
automatic	TITLE w/ C	ORGREF-tt-02	yes	no	no	0.1863	2.0185	0.2284	0.5000	0.2375	2.2697	0.3040	0.3875
automatic	TITLE w/ C	NAMAZU-tt-02	yes	no	no	0.1188	1.4345	0.1950	0.3875	0.1525	1.6568	0.2741	0.3000
automatic	DESC	GRACE-ds-02	yes	no	no	0.4163	5.1761	0.5833	0.1500	0.5013	5.6200	0.6452	0.1000
automatic	DESC	GRACE-ds-01	yes	no	no	0.3863	4.9528	0.5994	0.1375	0.4775	5.4242	0.6872	0.0750
automatic	DESC	DBLAB-ds-02	yes	no	no	0.3625	4.5517	0.5446	0.2500	0.4375	4.9293	0.6128	0.2000
automatic	DESC	DBLAB-ds-01	yes	no	no	0.3613	4.5465	0.5446	0.2500	0.4350	4.9197	0.6128	0.2000
automatic	DESC	SSTUT-ds-02	yes	no	no	0.2900	3.8716	0.4843	0.2625	0.3700	4.2896	0.5620	0.1875
automatic	DESC	SSTUT-ds-03	yes	no	yes	0.2688	3.6874	0.4681	0.2625	0.3575	4.1313	0.5462	0.2000
automatic	DESC	R2D2-ds-01	yes	no	no	0.2650	3.3843	0.4389	0.3125	0.3200	3.6953	0.5259	0.1750
automatic	DESC	TKB-ds-01	yes	no	no	0.2413	2.8832	0.3458	0.4250	0.3363	3.4132	0.4700	0.2750
automatic	DESC	K3100-ds-01	yes	no	yes	0.2338	2.9619	0.4002	0.3375	0.3025	3.3332	0.4814	0.2500
automatic	DESC	TKB-ds-02	yes	no	no	0.2275	2.8342	0.3618	0.3375	0.3100	3.3105	0.4818	0.2125
automatic	DESC	OKSAT-ds-01	yes	no	no	0.1688	2.0893	0.2841	0.5250	0.2088	2.3122	0.3309	0.4750
automatic	DESC	TKB-ds-03	yes	yes	no	0.1600	1.8526	0.2479	0.4750	0.2138	2.1748	0.3469	0.3750
automatic	DESC	SSTUT-ds-01	no	no	yes	0.1388	1.8667	0.2701	0.5500	0.1800	2.0961	0.3365	0.4500
automatic	DESC	TKB-ds-04	yes	yes	no	0.1088	1.3593	0.2241	0.5500	0.1550	1.5810	0.2741	0.4375
automatic	DESC	NAMAZU-ds-01	yes	no	no	0.0525	0.5803	0.0909	0.8000	0.0663	0.6674	0.1282	0.7375
automatic	DESC	K3100-ds-02	no	no	yes	0.0213	0.2158	0.0556	0.7125	0.0350	0.2924	0.0823	0.6500
interactive	TITLE, ALT1-3 w/ C	OKSAT-it-02	yes	no	no	0.2650	3.1823	0.3516	0.3625	0.3213	3.4797	0.4316	0.2750
interactive	TITLE, ALT1-3 w/ C	OKSAT-it-03	yes	no	no	0.2638	3.2691	0.4165	0.3125	0.3238	3.5922	0.4790	0.2500
interactive	DESC, BACK, TERM	TKB-it-01	yes	no	no	0.2500	3.0188	0.3593	0.4000	0.3113	3.3230	0.4036	0.3750
interactive	DESC, BACK, TERM	TKB-it-02	yes	no	no	0.2138	2.6064	0.3636	0.4000	0.2688	2.9269	0.4445	0.3125
mean						0.2416	3.0049	0.3784	0.3747	0.3071	3.3574	0.4601	0.2872

(In each query method and part of the topic used, the run ID codes are ranked in order of the precision at 10 document-level in DM_1 and RL_1.)

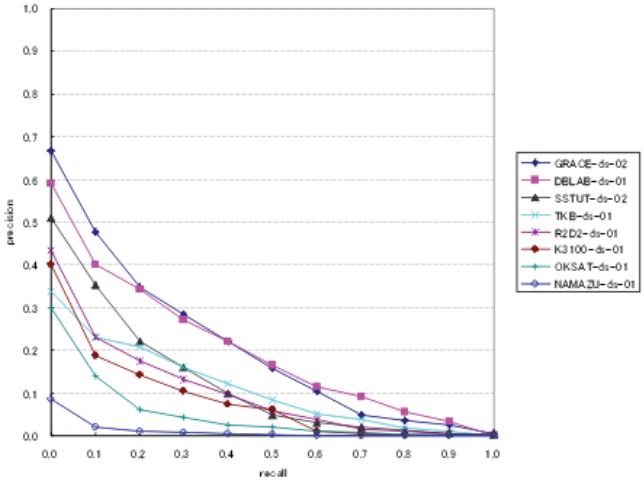
QueryMethod: Indicates 'automatic' or 'interactive'. 'Automatic' indicates a run without any human intervention during query processing and search; 'interactive' indicates a run other than 'automatic'.
TopicPart: Indicates the part of the topic used. 'w/ C' and 'w/o C' indicate if the system used Boolean operators that were specified as 'CASE' attribute or not, respectively.
RunID indicates the identification codes of the system run results, as shown in **Table 1**.
Cont: Indicates whether or not the system used textual contents of web documents for indexing.
Link: Indicates whether or not the system used link information in Web documents. The notation 'yes' indicates that the links and contents were used; 'no' indicates that only contents were used.
Anchor: Indicates whether or not the system used anchor text for indexing.
prec(10) indicates the precision at the 10-document level.
dcg(10) indicates the DCG value at the 10-document level.
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.



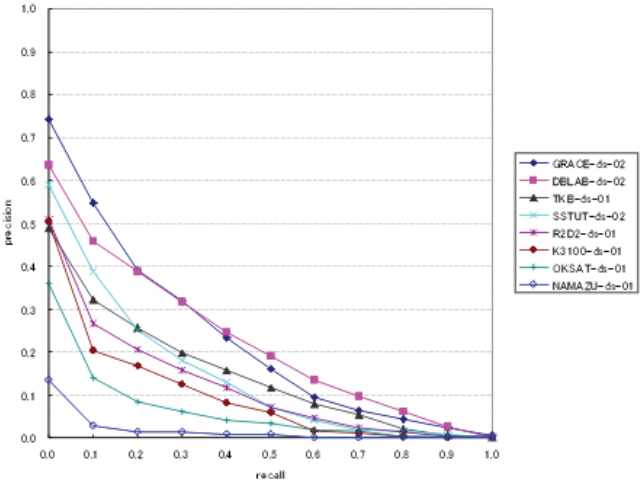
(a) Recall-precision curves for TITLE-only runs at RL_1 level



(b) Recall-precision curves for TITLE-only runs at RL_2 level

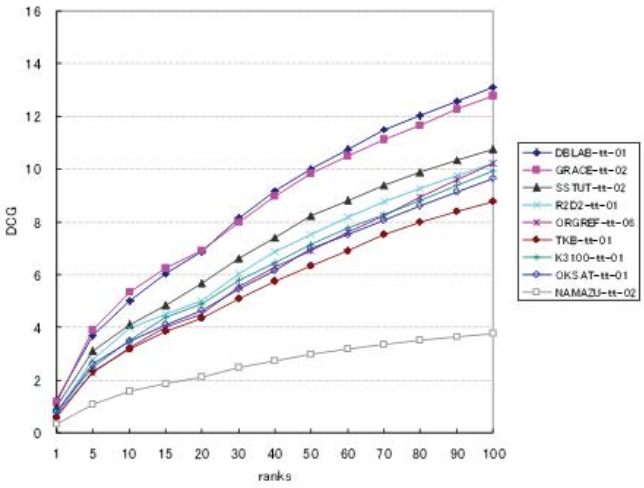


(c) Recall-precision curves for DESC-only runs at RL_1 level

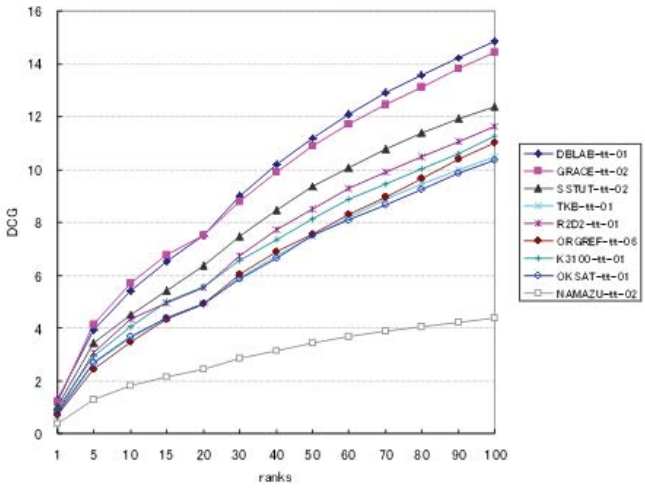


(d) Recall-precision curves for DESC-only runs at RL_2 level

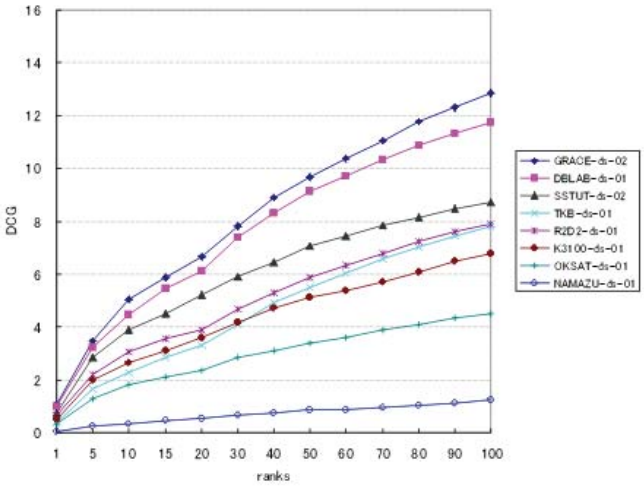
Figure 3. Recall-precision curves for the selected runs



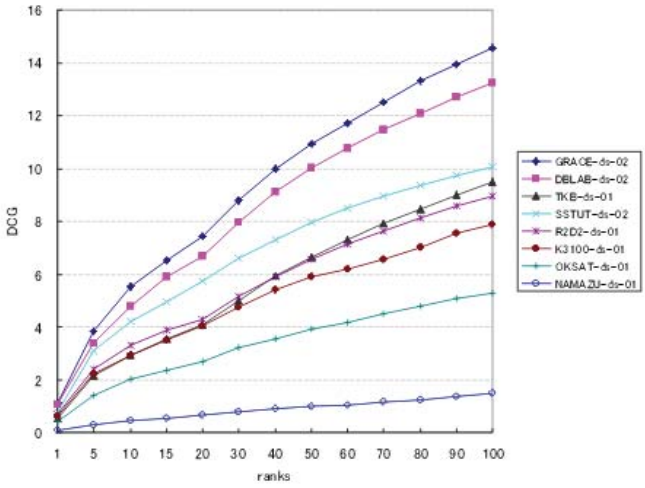
(a) DCG curves for TITLE-only runs at RL_1 level



(b) DCG curves for TITLE-only runs at RL_2 level



(c) DCG curves for DESC-only runs at RL_1 level



(d) DCG curves for DESC-only runs at RL_2 level

Figure 4. DCG curves for the selected runs